Large-scale integrating project (IP) project
FP7-ICT-2011-8

DILINET

Diverse **I**ndicators for **L**anguage **I**nfluence on the inter**NET**

PART B

Objective ICT-2011.4.4 - Intelligent Information Management
Target outcome b) Intelligent integrated systems that directly support decision making…
Keywords:
Horizontal Big Data; Decision Platform; Data Extraction; Language Indicators; Multilinguism; Web Search; Text Mining; Audio mining; Web measurement

Date of preparation: 17/01/2012
Version number: V40

Beneficiary no.	Beneficiary name	Short name	Country
1 (coordinator)	The European Research Consortium for Informatics and Mathematics	ERCIM	FR
2	World Network for Linguistic Diversity	MAAYA	CH
3	Istituto di Scienza e Tecnologie dell'Informazione	CNR	IT
4	Dialogic Innovation & Interaction	DIALOGIC	NL
5	Centre National de la Recherche Scientifique	CNRS	FR
6	Exalead, a branch of Dassault Systèmes	EXALEAD	FR
7	Universitat Pompeu Fabra	UPF	ES
8	Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.	FRAUNHOFER	DE
9	Dutch National Centre of Mathematics and Computer Science	CWI	NL
10	Fundación Redes y Desarrollo	FUNREDES	DO
11	Vocapia Research	VOCAPIA	FR
12	United Nations Educational, Scientific and Cultural Organization	UNESCO	FR
13	Nielsen Media Research GmbH	NIELSEN	DE

Proposal abstract

It is natural to see the Web as the best source of easily accessible, up-to-date horizontal Big Data for making decisions, reading the pulse of social situations. But what are we reading? The Web has mutated from a small initial collection of individually crafted pages written in English. No one now knows how to characterize the Web, its contours, its contents. Uniting specialists in Web indexing, mathematical sampling, graph analysis, text and speech processing, as well as international organisations implicated in world language policies, DILINET will create the tools and methods concerning language use that will allow informed decision making in multiple sensitive areas.

DILINET will create new methods of intelligent Web sampling, unbiased crawling, page cleaning, and language detection (both text and audio). International organisation partners will define language indicators from these sampling methods. Research partners will develop modelling techniques to characterize language use and web content, both text and video, providing quantification for the metrics. DILINET will also develop pro-active surveying of actual web use. Surveying partners will develop voluntary-use plug-ins to capture language usage in browsing situations and to vehicle short questionnaires to different populations. Through sampling and voluntary monitoring of web usage, DILINET will provide the first reliable measure of Web characteristics.

Providing new measures of information on the Web, useful both for science and government policy making, and new search strategies, DILINET applications will also demonstrate practical uses of this new information: measuring diversity in EU government websites, producing country specific Web indexes for underrepresented citizens, teaching materials for language policy makers, and targeted user campaigns for business usage.

DILINET will provide accurate sampling methods and trending tools directly exploitable for support decision making and situation awareness.

Table of contents

Figures and Tables

B 1. Concept and objectives, progress beyond state-of-the-art, S/T methodology and work plan

B 1.1 Concept and project objective(s)

The unstructured Web is the most difficult source of Big Data to exploit. The concept of DILINET is to use latest techniques in Web science, in mathematical sampling, and in language processing and modelling to characterize the Web, separating content from spam, identifying how information is spread over the web, producing replicable metrics and applications exploiting this characterization.

The most vital characteristic, and one of the hardest to measure, is language use: in terms of what information is offered in what language, in what countries, and how people actually use language to interact with the web.

DILINET enlists language specialist from scientific, industrial and international organisations to design measuring techniques and practical tools that will characterize real language use on the Web.

The objective of DILINET is to provide the first verifiable large scale characterization of the Web since the mid 2000s, opening new perspectives for Search Engines, helping the content industry better incorporate the linguistic factor. It will give to the content divide the role it deserves in the struggle against the Digital Divide and change the paradigm of measuring Information Society evolution.

B 1.1.1 The DILINET vision

The Web has changed. What started out as a collection of individually crafted pages (institutional sites, homepages) written principally in English for an international audience has changed over and over again, causing people to create iterative names such Web 2.0, Web 3.0.

The Web is a hodgepodge of informational pages, ecommerce stores, conversational and gaming platforms, spam, news feeds, aggregators, and contents which appears and disappears, or which is generated on demand. In the early 2000s people attempted to measure and characterize the web. Search engines proudly displayed how many pages they had crawled, and there were estimates of what percentage of the Web was covered by these search engines.

Since then counts have disappeared, as the appearance of automatically generated pages makes the web potentially infinite. Just like there is no last number, there is no finite set of web pages. Some sites will create a new page for every new URL. Attempts to measure and characterized have been mostly abandoned, and we no longer have a clear picture of the universe of the Web.

This lack of knowledge had impacts on society, economics, and culture. We do not know what new content is, and what is copied content or spam. We do not know how well a certain country is covered by a search engine. What percentage of real content pages is indexed? Since the number of independent web indexes has receded from dozen in the early 2000s to a handful today, there is an increasing lack of info-diversity, fewer ways to access information, fewer ways of being sure that voices can be heard.

Economic concerns are “where are rising markets for a given good or service”. Are new online use patterns arising on which new business can be developed? Are current business strategies still going in the right direction?

Cultural concerns have to do with preservation and expression of cultural identity. Can people within a given culture access information in their own language? Is this information available? Is it even indexed?

There are so many things we no longer know about the Web and so many uninformed decisions made in the new economy.

Language usage and presence are both critical variables, on all levels: societal, economic, and cultural. In the geography of the web languages are the ultimate frontiers which shall drive many decisions, yet it is also one of the hardest things to measure.

DILINET proposes to attack the problem of characterizing web content and usage for the third decade of the Web. We will develop techniques for sampling, validating, and characterizing content of the Web by probing the Web and search engines, and provide user-accepted add-ons to gather information on usage.

Output from this project will be reliable indicators, such as

- What are the languages of the Web?
- How much content appears in a given language?
- Is this content growing? How fast?
- What are the concepts used in a given language on the web?
- What is the frequency of content categories (e.g. politics, sports, sales ...) for different languages?
- What is the opinion on specific concepts, e.g. "Eurobonds", in different languages and countries?
- What are the temporal trends?
- What are the languages of medicine, art, science, and economics?
- What languages are spoken in web videos?
- What languages are used by people browsing, on social networks, in emails, in word processing?
- What are the connections between languages on the web?
- What is the lexicon of a language used on the Web?
- Which languages are growing?
- What countries are showing the greatest increase in content production? In what areas?
- Where are the dark zones, the areas of real content that are not indexed by Google, Yahoo, Bing, Baidu, Yandex, Exalead?
- What are the implications of the growing use of smart phones in terms of contents?

DILINET will provide the means of answering these questions which will allow informed decision making for many stakeholders.

B 1.1.2 Project Objectives

The goal of DILINET is to provide ***an intelligent integrated platform to support decision making*** concerning the use of the Web for knowledge mining. To realize this goal, we have identified a number of *strategic objectives*:

1. ***Better understanding of the real geography of the Web***. The Web is the primary source of unstructured Big Data. Though it is used for mining opinion, gathering citizen opinion, learning ontologies, and a wide variety of text mining tasks, its structure is poorly understood, like the Americas in the 16th century. DILINET will develop intelligent sampling techniques, web cleaning, and categorization techniques that will give us a clearer vision of what the web looks like. This is the topic of Work packages 5 and 8**.**

2. ***Better measurement of real language use on the Web.*** The language of the Web can be considered from two perspectives: offer and usage. Some of DILINET's partners have a wide experience in knowing what to measure to describe language usage. They will define indicators in Work package 4 which will guide the DILINET technical partners in which statistics to gather. Gathering will be from the static offer of pages (what a user could access through a web browser) *and* from sampling actual usage (what languages people use to interact with the web, what they read, etc). The first case is addressed in Work page 5 (sampling) and the second case forms the Work package 6 which will provide voluntarily-installed plug-ins that anonymously monitor users' language behaviour. From these two complementary sources, we will get a representative picture of Web use.

3. ***Better understanding of what is found on the Web.*** Large-scale sampling techniques will bring back cleaned and representative collections of what is found on the Web. From these samples, Work Package 8 will build up models, per language, per topic, and along other dimensions to extract the indicators of Work Package 4. Large scale language models and the tools needed to exploit them will be developed in Work Package 9. An added benefit of the DILINET project will be these models which can be used for other scientific work dealing with real language use.

4. ***Inclusion of audio and video analysis in addition to text.*** The world is becoming more and more multimedia, as video slowly replaces text. The growing place of video on the web cannot be ignored. DILINET will also measure language use in this newer media, developing new techniques for audio language identification, and improving speech to text results. This is the essential work of Work package 7.

5. ***Providing lasting tools for periodic measuring of web characteristics.*** The Web evolves at a tremendous rate. The purpose of DILINET is not to do a one-off measurement of the web but to perform repeated analyses to extract trends. It provides tools that can be used periodically, to follow the evolutions of the Web over time at the national and international level. Work Package 10 will develop solid architecture that will allow the other technical developments in DILINET to achieve this objective of creating an intelligent integrated platform to support future decision making.

6. ***Evaluating the value of the tools and platform directly with motivated stakeholders providing real life requirements and motivated decision oriented usages***. Beyond the research institutions involved, DILINET stands on the motivation of a combination of stakeholders (International Organizations, civil society and business entities) determined to offer concrete requirements and ready to evaluate and make use of the project results in their respective domains (including language policies, information society indicators, digital divide policies, Search Engine innovation, and Content Industry). Work Packages 4, 11 and 12 will cover the requirements, evaluation and applications aspects of the project as well as give insights for future impacts and roadmap for future research.

7. ***Offering hard data to allow the evaluation of usage trends and the impact of emerging technologies (as for example smart phones) on Internet usages and users behaviours***. Work Packages 5, 6 and 9, together and in synergy, will produce powerful tools to allow, directly or

through chronological series of data, to capture and evaluate the deep trends of evolution of the Internet and then permit decision makers, from business or from policies, to anticipate those trends and take the appropriate approaches.

Objective ICT-2011.4.4 Intelligent Information Management

b) Intelligent integrated systems that directly support decision making and situation awareness by dynamically integrating, correlating, fusing and analysing extremely large volumes of disparate data resources and streams. This includes (but is not restricted to) recognising complex events and patterns that are today difficult or impossible to detect, aggregating and mediating opinions or predictions, offering alternative conceptualisations, guaranteeing timeliness, completeness and correctness, integrating categorical and statistical analyses. Visual Analytics should equally integrate data analysis and visualization. The effectiveness of such solutions will be evaluated against the concrete requirements of relevant professionals and communities and tested on appropriately- sized user groups and extremely large data resources from the respective domains (including, but not limited to, finance, engineering, government, geospace, transport, urban management).

B 1.1.3 Project outcome

The outcome of the DILINET project will be:

- a platform for representative sampling and mining language dependant Big Data on the Web,
- a new method for analyzing page content, user opinion, and usage data over the web,
- new language indicators about actual language use on the web,
- new tools for mining multilingual audio streams,
- new language models of language use for detecting trends,
- characterization of which languages/regions are underrepresented in search engine indexes,
- advances in market analytics, speech recognition, and text analytics, particular for underrepresented languages,
- large corpora of representative Web content.

B 1.1.4 Relevance to the topics addressed by the call

In the table below we summarize how the envisaged project contributes to the objectives listed in the FP7 -ICT Work Programme 20011-2012, Objective ICT-2011.4.4 Intelligent Information Management, in particular for target outcome b)

Intelligent integrated systems that directly support decision making and situation awareness by dynamically integrating, correlating, fusing and analysing extremely large volumes of disparate data resources and streams. This includes (but is not restricted to) recognising complex events and patterns that are today difficult or impossible to detect, aggregating and mediating opinions or predictions, offering alternative conceptualisations, guaranteeing timeliness, completeness and correctness, integrating categorical and statistical analyses. Visual Analytics should equally integrate data analysis and visualization. The effectiveness of such solutions will be evaluated against the concrete requirements of relevant professionals and communities and tested on appropriately- sized user groups and extremely large data resources from the respective

domains (including, but not limited to, finance, engineering, government, geospace, transport, urban management).

We focus on expected outcomes and research themes as listed in the Call:

Expected Outcome	DILINET
Reinforced ability for a wide range of innovators to tap data infrastructures and to add value beyond the original purpose of the data through data analysis.	DILINET will clarify which parts of the Web can be profitably mined for extracting knowledge, providing tools for verifying accuracy, coverage, and identifying underexploited parts of the web than can be crawled and mined.
Reinforced ability to find, reuse and exploit data resources (collections, software components) created in one environment in very different, distant and unforeseen contexts.	DILINET will develop verifiable sampling techniques that can be applied by individual companies, by public or private institutions, to new information spaces on the web. As the web continues to change and morph, these generic techniques will allow us to continue to accurately measure web resources.
Value creation through extensive data collection and analysis.	DILINET will not only identify under-used parts of the Web, through its voluntary plug in surveys, DILINET will create a steady stream of actual web usage, providing data unknown beforehand. For example, several hundred terabytes of data will be repeatedly analyzed for trend extraction (see WP5). DILINET will also analysis how large portions of the web interact, between language groups and between countries.
Increased economic value of data resources or data analysis services through standards for validation, provenance, accountability, access and privacy control.	Through validation of the sampling techniques and verification of the metrics collected by DILINET, this project will provide a just and validated image of actual Web use, and the quality of Web data.
New scientific investigations enabled by large, interconnected data resources and attending infrastructure.	The unstructured part of the Web is the messiest and least understood part of the data continually generated on the web. DILINET with its investigations into intelligent sampling, language indicators, and knowledge modelling over the web will provide a new scientifically validated characterization of the Web which could open promising avenues for a new generation of Search Engines
Increased efficiency of organisations and better management of societal challenges (emergencies, planning ...) through more timely and better decision making.	By providing accurate language models, per country, per language, per topic, and by providing a new set of sampling techniques, DILINET will provide a base for accurately judging novelty and trends in the Web as well as practical tools to be used for the decision making

B 1.1.5 Timeliness of the proposed work

The Web is the largest source of constantly renewed unstructured data. Scientific progress over the past twenty years has made great strides in extracting semantically meaningful knowledge from textual sources, and new services, such as Data as a Service are being proposed to exploit the richness of the Web. Trend analysis, voice of the citizen, social concerns, e-Democracy, all these desires for better understanding involve exploiting the unstructured part of the Web. A project such as DILINET is timely, since there is a need for being able to validate that the information extracted from the web is both complete and accurate, providing a counterweight to the current trend of black-box monopolies for Web Search. The reaching of the last phase of internationalization of the Internet has implied new

challenges in terms of applications and governance of the Internet and the language variable has reached a singular importance as a consequence (Internationalized Domain Names is just the tip of the iceberg and the common claim of adding the new billion users untaps a new role for language and content in this endeavour) while absolutely no progress has been made in the capacity to manage with hard facts this now key variable.

B 1.2 Progress beyond the state-of-the-art

Beyond the state-of-the-art in current measures of web characteristics

The theme of characterizing the web and, in particular, measuring the linguistic diversity on the Internet, has remained for a long period the reserve of a small group of specialists. The interest on the linguistic characteristics of the Internet is at last gaining the recognition it deserves. Ahead of the development of internationalized domain names (IDN) and the relatively slow process under Action Line C8 of the WSIS Action Plan[1], interest in the theme is clearly growing, as much in international forums, such as the Internet Governance Forum (IGF), as in key international organisations in the field of the information society, such as UNESCO or ITU. The time is ripe then for the concept of developing language policies in the virtual world to take on added importance and become a reality. It is almost impossible to formulate policies in any field without a clear vision of the situation, and indicators enabling its development to be monitored and measured reasonably frequently. In the field of linguistic diversity on the Internet, UNESCO has published a report [1] on the current situation and evaluates its future prospects based on experience accumulated during twelve years of research in the subject area.

The situation it describes is in fact paradoxical and quite alarming. Until the late 1990s, it was marked by a lack of serious indicators which led to what amounted to disinformation exaggerating the place of English on the Web[2]. This period was followed by the preliminary work of a handful of pioneers, which pointed to some indicators, most of them focusing on the Web. However, at a time when interest in the theme is becoming universal, the work of these pioneers has ground to a halt or is experiencing difficulties as a result of developments in the Net and search engines[3]. Accordingly, there have been no productions to consult since 2007, when NUT[4] and FUNREDES / Union Latine, the two most visible actors in producing indicators published their most recent works.

The report published by UNESCO above mentioned gives an overview of the various approaches that have existed. Most of them had only one or two applications, which is incompatible with the need for sustainable indicators. Some results did have an impact, but either their validity did not resist scientific analysis, or they used overly imprecise figures provided by the search engines. Two methodologies stood out owing to their capacity to produce credible and useful indicators, the methods of LOP (managed by NUT) and of FUNREDES/ UNION LATINE.

[1] World Summit on the Information Society (WSIS) – http://itu.int/wsis

[2] As a matter of fact, confusion reigns as to the number of Internet users in a given language (the only rough figures available are provided by marketing companies) and the proportion of web pages in that language (a figure difficult to evaluate and which different methods have endeavoured to obtain).

[3] On the one hand, the size of the web makes the work of systematically browsing pages ("crawling") ever more problematic (we could say that is an infinite space), and on the other, partially as a result of the first point, search engines only index an increasingly insignificant percentage of the visible Web and the indications offered on the total number of occurrences of the words searched are no longer credible at all (even though some methods are based on them).

[4] NUT leads the *Language Observatory Project* — http://www.language-observatory.org/. See reference [4].

The methodology used by LOP [4], which is still in successful application[5], consists of systematically crawling all the pages of domains of countries to be studied and identifying their script in order to count the pages in a given language. Where a script is shared by a number of languages (as is the case of the Western languages), a language recognition algorithm is applied. The method reaches its limits though when it comes to focusing countries with a vast amount of pages, such as China and Korea and, for the same reasons, is not addressing generic domain names (such as .com, .net or .org). Furthermore, the trend in managing country code top level domains (ccTLD) is to allow its utilization by entities remote to the country, for business purposes, and this may provoke growing biases in the results.

The methodology used by FUNREDES/UNION LATINE, limited to a number of Western languages[6], is based on the selection of a vocabulary in these different languages with appropriate characteristics in terms of equivalence, range and cultural neutrality. The counting by search engines of the pages corresponding to this vocabulary enables their respective percentages to be compared with statistical tools. This has allowed the largest and more stable history of measurements since 1998, including interesting results by countries or for other Internet spaces. This approach is however no longer reliable since the counters offered by search engines cannot be trusted anymore. In addition, it is no longer feasible to extrapolate the results to the entire universe since the space indexed by the engines represents now a far smaller proportion of the total space[7] (and especially as linguistic bias has appeared, as a consequence, in the sample indexed). In order to work properly, the methodology should now take as its basis the direct crawling of Web pages, in the same way as the LOP methodology, but this brings us back to the previous problem.

Both methodologies focus on the Web space; the second has enabled incursions into different spaces (such as that of newsgroups), as long as there are search engines that can be used in these other spaces, and has been able to undertake some very preliminary work on measuring cultural diversity through characters and famous persons representative of cultures.

The bottleneck today for creating indicators for linguistic diversity in the Internet is directly (crawling) and indirectly (search engines) linked to the exponential growth of the Web and its size of hundred of billons of pages which makes sequential crawling too long a process. The very nature of the web has changed transforming it into a dynamic and infinite space and there is no more attempt to characterize such a huge and moving target. However, more than ever such data is required for informed decision making by many stakeholders from policy makers tackling the digital divide[8] to industrial players of the digital economy (especially those concerned by the content oriented segment) willing to establish reliable business cases for their ventures and understand the evolution of the markets.

In this context, it is urgent and indispensable to mobilize existing actors and encourage new ones to engage in ambitious, serious and collective research activities which could break the deadlock and expand the potential for building indicators for spaces other than the static Web and for other approaches than those involving a static vision of existing resources or maintaining a flat perception of content in absence of attempts to measure quality.

References:

[5] LOP after having measure languages in the Asian and African Web is conducting similar studies in Latin America and the Caribbean.

[6] Catalan, English, French, German, Italian, Portuguese, Spanish and Romanian.

[7] This percentage has fallen from 80 % to less than 30% in recent years.

[8] While historically the efforts to overcome the digital divide has been essentially focused on providing more physical access, a new trend is prone to arise, to focus the content divide which is linked to digital literacy and linguistic diversity and which figures provided by NUT and FUNREDES/UNION LATINE have shown it is an order of magnitude deeper than the access divide,

[1] D. Pimienta, D. Prado, Á. Blanco, *Twelve years of measuring linguistic diversity in the Internet: balance and perspectives*, UNESCO, 2009
Available online: http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

[2] ITU, *Measuring the Information Society: the ICT Development Index*, ISBN 92-61-12831-9, 2009

[3] J. Paolillo, D. Pimienta, D. Prado, et al. (2005), *Measuring Linguistic Diversity on the Internet*, UNESCO, 12/2005
Available online: http://unesdoc.unesco.org/images/0014/001421/142186e.pdf

[4] Suzuki I., Mikami Y., and al. (2002), "A Language and Character Set Determination Method Based on N-gram Statistics", *in ACM Transactions on Asian Language Information Processing*, Vol.1, No.3, September 2002, pp.270-279.

Beyond the state-of-the-art in Statistical sampling

In DILINET we overcome the current limitation in Statistical sampling by following three different, interrelated, directions: Web Crawling, Data Storage, and Repeated Sampling.
Regarding Web Crawling, DILINET has a particular characteristic. Differently from traditional search engines, DILINET's crawling must generate a collection of web pages, which can be considered as a weighted random sample of web pages. Only such a sample allows projecting the weighted averages of indicators to the whole web page population. Most algorithms are based on Markov Chains, which for connected and a periodic link networks converge to a unique stationary distribution. To enable and facilitate this type of sampling we will extend the methodology in the following directions. First we will derive efficient techniques to estimate the sampling weights, which are, for instance, inversely proportional to the PageRank. Second we will develop monitoring methods which indicate, if the sampling process has reached a steady state independent of the starting pages and if the sample size is large enough for the intended accuracy of estimates. Finally we will develop focused sampling approaches which oversample web pages with specific characteristics without impairing the randomness of the sample [cf. M. Kurant, M. Gjoka, C. T. Butts, A. Markopoulou 2011, Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks. SIGMETRICS 2011].
The second direction regards the **data repository**. Our repository will store documents in our collection in a compressed format that would guarantee both fast access to any required document and fast insertion of new documents. Our repository will be based on Lempel and Ziv 78 (LZ78) parsing as compression algorithm. This well-known compressor that has some interesting properties that we will exploit in order to design and implement fast queries and achieve compressed space. First of all, it is known that LZ78 achieves a good level of compression when applied to texts in natural language. Moreover, this level of compression tends to rapidly increase when, as in our case, it is used to encode very large collections. Another important feature is represented by the fact that with LZ78 we will be able to decompress any document in time proportional to its length (namely, it is not required the decompression of the whole collection). Finally, the insertion of a new document to the collection will again require time proportional the length of this new document. Also, DILINET will consider diverse data formats such as: graphs, multimedia, etc. We will extend the techniques studied for text-based data on these diverse data formats as well.
Regarding **Repeated Sampling**, to monitor the temporal evolution of the web we will repeat the sampling process in regular intervals. To detect changes while reducing variance we will revisit a part of the earlier collected domains and web pages. In part we will include new units to detect new developments. While current longitudinal web surveys [Eytan Adar, Jaime Teevan, Susan T. Dumais, Jonathan L. Elsas (2009): The Web Changes Everything: Understanding the Dynamics of Web Content, WSDM 09] collected some 50000 pages our sample will contain hundreds of Terabytes of web pages, which poses specific requirements to sampling algorithms, storage management and evaluation.

Beyond the state-of-the-art in Polling

Current user-centric measurements are limited to either recording the URL's people are visiting and/or the search terms they are using. Language identification can be applied to the search terms but is fairly limited in the case of URL's. In DILINET, we will link the URLs with an extensive database of websites that have already been characterized in terms of language. Another very important limitation of current user-centric measurements is that they miss all off line behavior. By extending the queries to locally stored files, our measurement client is able to detect the language type of off line text documents as well. To safeguard privacy, an innovative heuristic will be used that built on the off line sampling heuristic from Zhang and Van Alstyne[9]. The heuristic enables the recognition of specific words (which function as language markers) without revealing the actual semantic content of the personal documents. A final shift in the current technology frontier is the combination of the online client with an online survey module. Focused survey questions can be triggered by specific events (that is, online usage patterns). Although such event-driven online surveys applications are already commercially available, with is new here is that the events are based on actual language use, not on basic indicators such as a specific URL that is being visited.

Beyond the state-of-the-art in language processing

There are two main approaches used by language recognition systems: acoustic and phonotactic. Acoustic approaches rely on the acoustic parameters derived directly from the speech signal, and modelled with discriminatively trained Gaussian Mixture Models (Burget, 2006) or with Support Vector Machine-GMM Super Vector (Torres-Carrasquillo, 2008). The performance of acoustic approaches has improved significantly using a variety of channel compensation techniques (Kenny 2007, Castaldo 2007). Phonotactic approaches rely on the assumption that the sequences of phonemes, that is how the sounds follow on another in words and sentences is language specific (Harper,2008). This means that even if two languages share the same set of phonemes, their phonotactic characteristics are different. Phonotactic approaches typically use one (Phone Recognizer followed by Language Modelling (PRLM) approach) or multiple (Parallel Phone Recognizers followed by Language Modelling (PPRLM) approach) well trained phone recognizer(s) (Zissman, 1996) to derive phone n-gram statistics. These phonotactic characteristics are then used to estimate models for language recognition. Today's best performing language recognition systems combine both phonotactic and acoustic sub-systems (BenZeghiba 2009). Several techniques that have been widely adopted in speech recognition have not yet been investigated for language identification. These include discriminant training for the component acoustic phone models; methods to improve the quality of target language phonotactic models such as optimization of decoding parameters and intelligent selection of phone contexts; and improved decoding making use of multiple hypotheses and automatic learning techniques.

For the most part Language Identification systems have been developed for the classification of telephone speech, and evaluated in regular benchmarks run by the US National Institute of Science and Technology (NIST) www.itl.nist.gov/iad/mig//tests/lre. In WP7 of DILINET, state-of-the-art techniques will be adapted and improve so that they can be successfully applied to the heterogeneous audio data found on the web.

References:

L. Burget, P. Matejka and J. Cernocky, Discriminative Training Techniques for Acoustic Language Identification, Proceedings of ICASSP'06, Vol. 1, pp. 197-200, Toulouse, 2006

[9] Zhang, J. and M. Van Alstyne (2003). EmailNet: A System for Automatically Mining Social Networks from Organisational Email Communications. North American Association for Computational Social and Organisational Science (NAACSOS), Pittsburgh.

P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, Joint Factor Analysis versus Eigenchannels for speaker recognition, IEEE Trans. Audio, Speech and Language Processing, 15(4), pp. 1435-1447, May 2007.

F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair, Compensation of Nuisance Factors for Speaker and Language Recognition, IEEE Trans. Audio, Speech and Language Processing, Vol 15(7), pp. 1969-1978, 2007.

P.A. Torres-Carrasquillo, E. Singer, W.M. Campbell, T. Gleason, A. McCree, D. A. Reynolds, F. Richardson, W. Shen, D.E. Sturim, The MITLL NIST LRE 2007 Language Recognition system, Interspeech'08, Brisbane, pp. 719-722.

M. P. Harper and M. Maxwell, Spoken Language Characterization, Springer Handbook of Speech Processing, Chap. 40, pp. 797-807, 2008.

M. A. Zissman, Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, IEEE Trans. Speech and Audio Processing, 4(1):31-44, 1996.

M. F. BenZeghiba, J. L. Gauvain and L. Lamel Language Score Calibration using Adapted Gaussian Back-end, Interspeech'09, pp. 2191-2194, Brighton.

Beyond the state-of-the-art in language indicators as input to language policy

Of the 6000 or 7000 languages in the world, there would be between 150 and 200 which have corresponding institutions that assure their protection, development or equipment and about 500[10] which are localized in the Internet (e.g. have a specific coding scheme and have a written presence in cyberspace) although much more have audio resources and some 1000 languages show some trace of presence[11]. The proper development of a language is not only the result of the willingness and ability of an institution to make it suitable for any use; it is often the desire of individuals, teachers and small associative structures working in the field.
Whether it is a solid institution or smaller initiatives, there is no normative or science or praxis model in terms of language policies. While many books and many conferences are dedicated to it (some explicitly, but more often implicitly) language policy (or language planning) is a subject of study only since a few decades. Furthermore, because of the disparity of sociolinguistic situations, linguistic policies studies can effectively guide practitioners only to the best-studied languages (including Asian and some European languages).
Thus, it is difficult to find two consistent definitions of language policy for the country or territory where the policy is applied. The term "language policy" covers a fairly heterogeneous set of activities ranging from promotion policies, protection or revitalization of one or more languages, to policy for eradicating them. Thus, we find very elaborate and explicit policies designed to revitalize a language becoming minorized or rarely used (Quebec, Catalonia, Israel), other for the political development of indigenous languages (Mexico, Bolivia, Paraguay, Mali, Benin, etc. .), equity policies of citizens vis-à-vis their states (Switzerland, Luxembourg, Aruba, etc..), but also linguistic hegemony policies giving priority to the state language (implicit as in United States or, explicit, banning the use of others languages, as in Italy during the Fascist period or in Spain during Franco period).

[10] The figure of around 500 comes from the number of languages identified by Unicode having a digital representation (http://unicode.org/repos/cldr-tmp/trunk/diff/supplemental/languages_and_scripts.html); from those an estimated 300 are effectively used in the Net (the highest figure come from Wikipedia with 269 different languages represented (http://stats.wikimedia.org/EN/Sitemap.htm)
[11] http://borel.slu.edu/crubadan/stadas.html

Leclerc, in his page dedicated to the language policies of CIRAL[12], proposes a typology differentiating between State policies:

- **Assimilation**, using actions to accelerate the liquidation of minority status or language groups.
- **No-action,** choosing the path of laissez-faire and let evolve normally the ratio of forces
- **Enhancement of the official language**, promoting one language in the political, legal, social, economic, etc.
- **Sectorial**, by adopting legislation more or less developed in one, two or three areas in the use of minority languages and immigrants
- **Differential legal status**, claiming harmonize linguistic cohabitation and legal equality for all. The majority of the territory has all the language rights; minority groups have rights in key sectors such as government services, justice, schools, and media.
- **Bilingual (or trilingual) recognizing the equality of two or more languages**, usually in legal or in the relationship between citizens with the State
- **Strategic multilingualism**, considering a language as complementary to one another and endeavour to operate in a positive way all the linguistic resources of the country
- **Language internationalization**, when states, generally former colonial powers exert their supremacy in terms of the linguistic code beyond their political boundaries
- **Mixed**, namely by the simultaneous practice of different types of intervention.

A language policy, for example, may intervene for the adoption of a system of writing or grammar and spelling, by creating vocabulary, by determining the status of one or more languages (official, language of administration and justice, language teaching, regional or national language, etc..), or for its teaching at the international level.
"Language policy" involves any decision concerning the use, expansion, corpus, status or education of one or more languages spoken in the territories concerned or having a requirement to be learned.
The indicators are the fundamental guiding tool for language policies, whether they witness a linguistic hegemony or on the contrary, whether they promote equitable rights for citizens, allowing them to participate fully in society through the use of their respective language.

As a matters of fact, although many national or international bodies (UNESCO, OIF, Union Latine, SIL, University of Laval, British Council, CPLP, etc..) offer statistics, surveys, studies, and other types of indicative data and try to find parameters more or less stable for the observation of one or more languages, indicators derived from them differ from one institute to other providing glaring inconsistencies and allowing communication buzz which in turn only create greater uncertainty.
From the simple question of how many languages exists in the world (which raises the most difficult issues about the border between languages and dialects, for example), to the question of how to measure the number of speakers (particularly in terms of their ability or level of mastery of the language), many other questions are emerging, such as the size of the corpus available, or the actual use in daily life (education, administration, health, science…)

New parameters are fed into the arguments for giving more room for one language over others or to balance the presence among languages, how much "weights" a language or how much "worth" a language. Various works done by Grin[13], Graddol[14], López Delgado[15] and Esperança[16] relates to the value

[12] http://www.tlfq.ulaval.ca/axl/monde/index_politique-lng.htm, consulté en décembre 2011.

[13] Grin, F., Compétences et récompenses. La valeur des langues en Suisse, Fribourg, Éditions Universitaires Fribourg, 1999 and Grin, F., « English as economic value: facts and fallacies », World Englishes, n°20, 2001

[14] GRADDOL. D. English Next. Why Global English may Mean the End of "English as a Foreign Language". The British Council & The British Company (UK) Ltd. 2006 (nlle éd.. 2007). En ligne sur <http://www.britishcouncil.org/learning-research-english-next.pdf>

of a language (i.e. English, French, Spanish and Portuguese) as well as those conducted by Calvet[17] [18] and others[19] on the weight of language, trying to give verifiable parameters on the opportunity to learn or teach a language, the need to promote the presence of a language into societal sectors where it is less present, or simply to position themselves on the labor or product markets.
The observation of the evolution of languages in cyberspace is no exception to this rule, and can highlight the existing gaps in this area. It is precisely the emergence of the Internet that asks new "opportunity" and "issue" of this new medium for languages. In fact, cyberspace is a real challenge for any language that is faced with providing more information than others, in order to avoid its speakers to phase out by adopting what other language that would be in their eyes more "prestigious". But it is also an opportunity because this medium allows an easier and less expensive capacity of expression compared with traditional media (including print edition or radio broadcasting) and thus, become the gateway to the resurgence of this language.
However, only people aware of this duality opportunity/challenge will help their languages flourish; others will wait to the certain death of their languages in a globalized world where the digital world gradually impose its rules worldwide.
DILINET, while proposing the creation of indicators on the place of languages in cyberspace to better measure the available contents will also put in parallel the existing information about weights on the real place of language in society, paving the way for creating universal indicators for language planning criteria and decision making. Finally, it will aim to educate all stakeholders (institutions, associations, educational, etc..) that can help all the languages of the world to make them visible and productive throughout the digital world.

Beyond the state-of-the-art in Automatic speech recognition

Automatic speech recognition is concerned with converting the speech waveform, an acoustic signal, into a sequence of words. Today's best performing approaches are based on a statistical modelling of the speech signal. The CNRS and VOCAPIA Research have collaboratively developed speech-to-text transcription systems for over 15 languages. Today's transcription systems are typically trained on huge heterogeneous audio and text corpora. In the DILINET project the CNRS and VR will apply unsupervised methods to train acoustic models for a number of additional languages (selected in collaboration with the user partners) to increase the language coverage.

Beyond the state-of-the-art in Text Categorization & Opinion Mining

Category recognition aims at classifying web documents to a hierarchy of predefined content categories[20] Opinion mining (also called sentiment analysis) focuses on the automatic identification of subjective expressions that describe people's sentiments or feelings toward entities, events and their properties[21], e.g. as positive, neutral, negative. State-of-the-art approaches perform category recognition or opinion mining for at most a few languages. DILINET poses a completely new challenge as it requires content classification and opinion mining for very many languages such that the content

[15] García Delgado, J.L. et al, Economía del español. Una introducción (2ª edición ampliada), Madrid, Editorial Ariel, 2008.
[16] Esperança, José Paulo**.** http://www.portalingua.info/fr/actualites/article/valor-economica-portugues/
[17] Calvet, Louis-Jean (2002). Le marché aux langues. París: Plon.
[18] See Poids des langues in Portalingua : http://www.portalingua.info/fr/poids-des-langues/
[19] Maurais, Jacques y Morris, Michael A. (eds.) (2003). Languages in a globalising world. Cambridge: Cambridge University Press; Wallraf, Barbara. "What global language". The Atlantic Monthly 286: 5 (2.000): pp. 52-66. ; Weber, George. "Top Languages". Language Today. Diciembre (1997): 12-18
[20] Carlos N. Silla Jr. & Alex A. Freitas 2011: A Survey of Hierarchical Classification Across Different Application Domains. Data Mining and Knowledge Discovery, 22(1-2):31-72, January 2011.
[21] Bing Liu (2010): Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, 2nd edition.

categories and opinion values have to be identical across all languages. Advanced algorithms need annotated training data which are not available except for a few languages. To improve supervised learning we first derive multilingual topic models using available multilingual resources like dictionaries, automatic translation, Wikipedia, or linked open data[22].

Content

As a hierarchy of categories we will select a resource that provides training data in at least a number of languages (like Wikipedia). We will enhance this training data by cross-lingual projections of training data from major languages, e.g. by automatic translation. For content categorization we will extend available algorithms[23] using topic model features and exploiting cross-lingual projections of training data.

Concerning content search in audio, speech processing technology will be exploited to provide the possibility to search and select audio and video documents directly from their contents. This is achieved by using automatic transcription and topic/content detection methods to annotate the unstructured data sampled from the web. The amount of audio and video data on the web has increased dramatically: YouTube estimates that 48 hours of content are uploaded every minute. In light of the enormous volume of data it is evident that automatic tools must be used to enhance the accessibility, usability and traceability of web content enhanced.

Opinion/Sentiment

Simple multilingual opinion mining approaches use a dictionary of words expressing positive and negative opinions and count their number[24]. Again a usually better alternative is to construct an annotated training data set by cross-lingual projections from a major language[25]. Enhanced topic models may represent the connection of topics to sentiment ratings[26]. By using features drawn from multiple languages even a better opinion mining performance can be reached than for single languages[27].

For audio data, the opinion analysis will be based on statistical models trained on pre-selected data associated with the transcription. Based on the output of the speech-to-text system, linguistic analysis will allow detection of opinion-carrying words. A paralinguistic analysis based on measures in the speech signal, such as the speaking rate, energy, pitch, as well as speech fluency (hesitations, pauses) frequency, will provide an indication of the observable opinion or sentiment of the speech. The information carried in the audio and in the transcription will be fused.

[22] Jagadeesh Jagarlamudi & Hal Daume III: Extracting Multilingual Topics from Unaligned Comparable Corpora. ECIR 2010.

[23] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, Zheng Chen (2011): Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia. WSDM 2011: 375-384.

[24] C. Banea, R. Mihalcea, and J. Wiebe (2008): A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In LREC 2008. Zheng Lin, Songbo Tan, Xueqi Cheng (2011): Language-independent Sentiment Classification Using Three Common Words. CIM2011.

[25] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan (2008): Multilingual subjectivity analysis using machine translation. EMNLP 2008.

[26] Jordan Boyd-Graber, Philip Resnik (2010): Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation. EMNLP 2010.

[27] Carmen Banea, Rada Mihalcea (2010): Multilingual Subjectivity: Are More Languages Better? COLING 2010.

B 1.3 S/T methodology and associated work plan

B 1.3.1 Overall strategy and general description

The work is organized in a number of technical work packages (WP5-9), with the other work packages focusing on legal, societal and policy issues concerning measuring language use (WP4), on integration (WP10), on dissemination and exploitation (WP3), on applications (WP11) and on evaluation (WP12), and two work packages dedicated to project management (WP1) and scientific coordination (WP2).

	Title		Lead
WP 1	Project Management	1	ERCIM
WP 2	Scientific Coordination	6	EXALEAD
WP 3	Dissemination & Exploitation	2	MAAYA
WP 4	Societal Issues	2	MAAYA
WP 5	Smart Sampling of Large-Scale Data	3	CNR
WP 6	User Centered Measurement	4	DIALOGIC
WP 7	Language Indicators in Non-Text Media	5	CNRS
WP 8	Validity & Analysis	6	EXALEAD
WP 9	Data & Knowledge Representation	9	CWI
WP 10	System Development and Integration	8	FRAUNHOFER
WP 11	Applications	7	NIELSEN
WP12	Results Assessment & Evaluation	10	FUNREDES

Figure 1 – Overview of work packages

B 1.3.2 Timing of work packages and their components

B 1.3.2.1 Project planning

DILINET will implement its work plan over three years, with major deliveries at M12, M24, and M36.

B 1.3.2.2 Overall work description

DILINET work can be divided into three parts: technical work on measuring the Web, intelligent sampling, and extracting useful data and metrics (WP5, WP6, WP7, WP8, WP 9, WP10, WP12); applicative work using this data (WP4, WP11); and dissemination and management (WP1, WP2, WP3).

Here is a description of the non-management work packages.

1.3.2.2.1 *Technical work package overview*

WP5: Smart Sampling of Large-Scale Data

Summary:

We develop techniques for deciding how to produce a large representative sample of the Web, design and implement web crawlers that deploy these sampling techniques, and store the results in large but efficiently accessible data repositories. In our case study of language use on the web, this work package provides us with an image of what the web offers in terms of information.

Data is pervasive in everyday people life and getting to better understand aspects of such huge amount of data is of paramount importance as a driver of decisions. As “no good deed goes unpunished” we cannot think of unravelling the hidden content and to measure indicators on the whole dataset. Rather, we have to consider the best combinations of crawling and sampling possible in order to cover as much as possible the data space we have to analyze with the smallest estimation error possible. Crawling, indeed, is already a sampling mechanism. It is, indeed, a very careful sampling mechanism, as a crawler has also to be “polite” with respect to the servers that are contacted. A web crawler is a software system designed to accomplish such a task. In the design and development of such a complex object many goals come into play such as, on one hand, efficiency with respect to computational and networking resources on the harvesting side, and, on the other hand, politeness to who is hosting the harvested content and significance of collected data with respect to its usage. Moreover, the dramatic evolution of the web has made such goals fast moving targets, not only because of the increasing size of the web itself (estimated today in dozens of billions of pages, amounting in petabytes of data), but also due to the ever increasing diffusion of "web spam" that pushes the need of adversarial approaches to fight such phenomena.

Traditional web crawlers were concerned with harvesting “high quality” pages. In this regard, high quality is considered as an indicator of interestingness. As it is impossible to crawl the whole web DILINET has to collect a representative sample of web pages. Ideally it should be composed of a large number of pages selected randomly from the set of all web pages. There is, however, no list of all web pages, and DILINET has to follow hyperlinks of web pages to collect a sample. To arrive at a representative sample this has to be done in a random fashion. However, web pages with many in-links have a much larger probability to enter the sample. This probability corresponds to the PageRank of that web page [Eda Baykan, Monika Rauch Henzinger, Stefan F. Keller, Sebastian De Castelberg, Markus Kinzler: A Comparison of Techniques for Sampling Web Pages. STACS 2009: 13-30]. Besides collecting a random sample of web pages the sampling process therefore requires the approximate estimation of the PageRank of the collected pages which then is used for the weighted aggregation of the diverse indicators, e.g., language diversity, languages per country, content categories per language, etc.

The size of the sample determines the accuracy of estimates, which depend on the size of the subgroups considered. Very many small languages cover much less than 0.1% of all web pages[28]. If we consider a small language covering 0,01% of all web pages and we need about 10000 documents to represent that language in terms of content topics, opinions, etc., then a full sample would require 10000/0,0001 = 100 million pages. We may assume a fraction of 5% videos/audio (10 MB/video) and 95% Text (200k/document). This gives about 5 million audio/video documents (50 Terabytes) and 95 million web documents (19 Terabytes) yielding 69 TB for a sample. As we have to estimate the number of in-links for each document to get the sampling weight a multiple number of pages have to be visited and analyzed during the sampling process. To extract temporal trends we repeat the whole sampling process in regular intervals again multiplying the size of collected data. On the other hand focussed sampling, which tries to access web pages having specific features with a higher probability, may reduce this number. Note, however, that we always require reliable estimates of the fraction of the considered subpopulations and therefore can only reduce the above sizes by a moderate degree. In summary we estimate that the construction of a single sample requires downloading several 100 TB of web pages.

[28] http://w3techs.com/technologies/overview/content_language/all

The vast amount of raw data produced by the sampling/crawling strategy in DILINET must be stored to be processed by upper levels. In this WP, then, we also deal with a storage module that needs to be ready to accommodate various petabytes of multimedia and heterogeneous data produced by the crawlers. Obviously, we cannot leverage on single machines to accomplish to this task and we will, thus, resort to use distributed processing techniques enabling to scale up with the data growth expected. At a first sight, the huge amount of data we envision to use seems to be in contrast to the fact that we want to exploit sampling. A more careful analysis, though, will make this point valid.

According to Nielson-Online currently there are more than 1,733,993,741 internet users generating a whopping number of 1.8 zettabytes being created and replicated (as in copied to DVDs and shared in the cloud) this year alone — a number that doubles every two years, according to a recent study by IDC and EMC. But how much is that, really? Not only is data itself ethereal and hard to visualize, but the numbers are so gargantuan that they quickly become too abstract to grasp. Those 1.8 zettabytes of data, for example, would require 57.5 billion 32 GB iPads to store. How much is that? About $34.4 trillion worth. That's equivalent to the GDP of the United States, Japan, China, Germany, France, the United Kingdom and Italy combined. And that's how much data we'll create and store just this year. Therefore, sampling will help in not drowning in such an amount of data, and at least to be able to reduce it to up than six orders of magnitude.

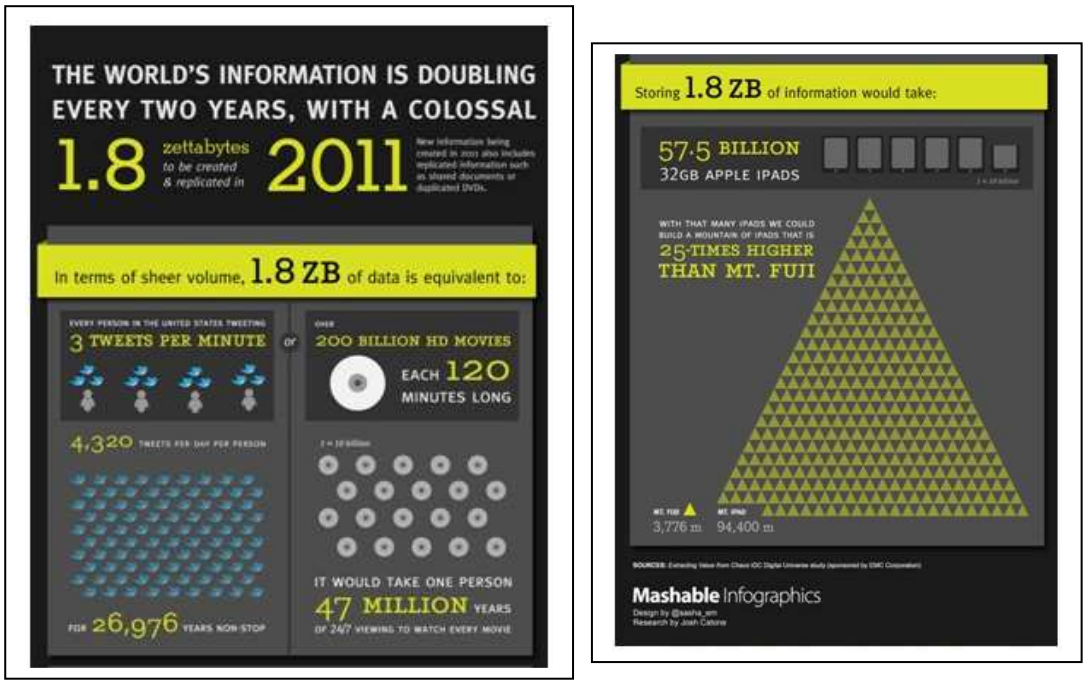


Figure 2 - The World's Information Production per Year

In these years, several open source solutions have been developed to deal with big data. From the most famous Hadoop (an open source MapReduce solution) to Pig, a data flow analysis tool running on top of Hadoop. In the spectrum, though, there exist many other solutions that go under the general definition of vertical databases, originated from the NoSQL movement. In DILINET we plan to use and extend such solutions in order to manage the amount of data we have.

For WP5 we distinguish three different tasks:

1. Task 5.1 is the task devoted to the Large scale data storage. This task designs and deploys the large-scale repository for the data-samples collected by the intelligent crawlers built in Task 5.3 and the indicators specified by WP6 and WP7. Data sources of interest for DILINET are

heterogeneous (i.e., they contain different kinds of textual and multimedia data). Moreover, the different modules in the DILINET platform that analyze data, extract and exploit the different Internet indicators designed, must access data locally and very efficiently. Therefore, a scalable and high-performance storage module is strongly needed. The storage module must support efficiently scan operations over large and compressed large-scale datasets. Also concurrent accesses to the data stored must be supported efficiently in order to allow multiple instances of DILINET analysis applications to be run concurrently. Furthermore, the nature of the data, in the areas relevant to DILINET, poses a lot of questions that need to be addressed from a storage point of view. For example, a lot of interesting data is represented in the form of a tree or a graph, rather than in relational form. Moreover, the operations on such huge graph are sometimes very complex. Thus, there is a need to explore again all the traditional research questions, such as fast access, query optimization and concurrent operations, in order to develop techniques that take full advantage of the recent advances in compression algorithms, and are suitable to perform complex operations on very large data samples in a scalable way. The purpose of this task is to design such a large-scale and high-performance infrastructure for storing and accessing the various kinds of DILINET data: raw textual data; indexes built to speed-up data access; knowledge, expressed in the indicators as defined in WP6 and WP7.

2. Task 5.2 will develop unbiased and efficient sampling strategies for web pages appropriate for the accurate estimation of statistical indicators of the web. First the target population has to be defined by the partners, with special considerations for cases like dynamic web pages. The hyperlinks of web pages form a graph which is directed and only outgoing links are easily discovered. To arrive at a random sample of web pages with known properties a random walk following randomly selected links has to be performed. The resulting web pages have to be reweighted, e.g. by their PageRank, as pages with many links are collected more often. Existing sampling approaches will be analyzed and new algorithms will be developed. The main emphasis will be on unbiasedness and efficiency, e.g. fast coverage of the whole sampling space. An additional line of work is the development of unbiased focused sampling approaches which oversample specific subpopulations. They may be triggered in an adaptive fashion, e.g. if the current accuracy of specific indicators is low. Prototypes of the algorithms will be implemented and evaluated on synthetic and real data.
3. Task 5.3 deals with Intelligent Data Gathering. Differently from traditional web crawling, in DILINET we do not need to maintain huge document corpora in a repository for future use. On the other hand, DILINET performs large-scale crawling and large-scale data analysis on the data samples retrieved in order to estimate the indicators of WP6-WP9 with the required precision. In this task different solutions will be investigated aimed at pushing forward the limits (and backward the costs) of distributed web crawling. For example volunteer computing will be investigated as a means of granting the huge network and computational bandwidth needed to actually download and analyze large data samples. As another example, the availability and data persistence properties of cloud storage services could be exploited to provide a convenient solution to the problem of efficiently orchestrating the instances of focused crawlers running on the volunteer computing platform. The main challenges we have to face in this task, anyway, are very interesting. The first very important challenge is how to trade-off between the qualities of the sample collected and the cost and the time needed to collect it. An open issue is also the choice of the most efficient way to implement unbiased and focused crawlers, and the best data structures and access methods to deploy and orchestrate them.

Prototypes of the sampling algorithms will be implemented and evaluated on synthetic and real data. We will measure the difference is between indicators estimated from a sample and the true figures in the population. In addition we will analyze if error intervals for estimates are realistic. In addition the techniques and tools developed in WP5 will be validated through the methods specified in WP12 and the usability of the data storage module will be assessed and refined through the project duration by the specification and analysis done in WP10 and WP11.

WP6: User-centric measurement

Summary:
In this work package, we produce tools for determining how users use languages on the web. We design a plug-in that volunteer users will download to their devices. This plugin will perform two duties: measure actual language use and feed survey questions to the user from time to time. Anonymized usage information will be returned from this plug-in, respecting the privacy restrictions determined in WP4. In our language use case, this work package reveals how language is actually being used by the user sample, in various applications and at various moments through the day.

This WP measures language behavior at the most detailed level and at the point of origin: the individual user. The measurement uses a sophisticated piece of software that is being installed at one or more of the devices that is being used by the end user. This includes desk top PC's and laptops and mobile devices (iPhone and Android smartphones and tablets). The measurement instrument consists of two parts that can be used independently but can also be combined: an online survey module that can be used to directly forward questions to the user, and an automated robot that classifies which language is being used online and offline.
The automated robot is a small script that automatically identifies the languages that are being used by the end user on that particular device, in various settings (e.g., online use of social networks, use of chat and VoIP, offline use of text processing and so on). It basically covers any text that is being typed and temporarily stored on the device. Theoretically, the optimum solution would be to measure input at the lowest OSI-level, that is, key strokes. However because some types of known malware use a similar technology (kernel- and API-based keylogging) our client would be blocked by antivirus software (e.g., Windows Defender). We therefore use temporary stored content instead (cache) as a data source. Users have full control over the privacy settings. For the online tracking behaviour, for instance, they can block specific websites for being tracked, or temporarily pause tracking.

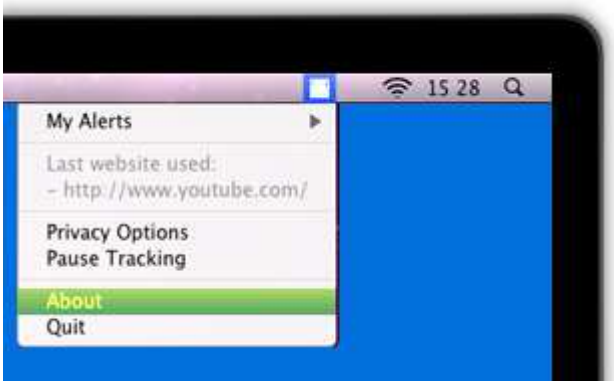


Figure 3 - Dummy screenshot of the online module

A major technical challenge is to keep the client light. This is especially important for mobile devices, where the use of memory and processor capacity should be kept as low as possible. At the same time, the identification of the language requires the use of a large number of extensive libraries. Subsequently the actual identification of the language should be done at the server side, not at the client side. This would however require sending the original input files (which might contain personal data) to the server. This does not comply with our strict standards of privacy by design that states that under no circumstances any sensitive data should leave the client. To still be able to identify the language without disclosing the actual content of the input file we use a heuristic similar to the one that has been used by Zhang & Van Alstyne (2003) for the anonymous analysis of email communications. The heuristic uses one-way hash functions that permit comparisons of similar tokens but not semantic interpretation of

content. The method only takes a limited number of random words from the input file and sends these over an encrypted connection to the server. Once the algorithm at the server has identified the language, the string of words is being erased.
Although the client only collects a modest amount of data for each unique language action of the end-user (a string of characters), the total amount of data generated is still very big due the very high frequency of the measurements. In previous pilot studies, with a relatively small panel size of 10,000 active users, the client already generated 250 megabytes of data on a daily base, or 7.5 gigabytes per month (Brennenraedts & te Velde, 2011). Note that this measurement campaign only covered a small part of the online behavior, namely the URL's in the browser cache. The proposed measurements in this project have a much wider scope (e.g., include all online behavior and also offline behavior) and contain much more details. Hence the input files and processed files will be several magnitudes (50-150 times) bigger than in the previous pilot. We estimate the total monthly amount of data that will be generated between 0.4 and 1.1 terabytes.
The second part of the client is an online survey module. This is a more traditional measurement instrument. Yet this version has two innovative twists. First, the anonymity of the respondents is strictly guaranteed by using hash functions to link the recipient to her or his background variables. This means that we can select specific target groups and sending the respondents within those target groups tailored questions without ever knowing their identity. To further ensure (k-)anonymity, the front-end of the survey tool automatically tests whether the combination of the background variables does not lead to too small subgroups where unique individuals could in theory be identified (Sweeney, 2002). Secondly, tailored questions cannot only be initiated by researchers but also by the client itself. This method used the first part of the client, the automated agent, to trigger specific questions. The questions and triggers are programmed in advance by the researchers but triggered by specific events in the automated agent. This means that we can immediately follow-up specific patterns of language behavior by specific groups of users with specific questions (e.g., to clarify the behavior and/or asking more detailed information or types of variables that cannot be covered by non-intrusive automated data collection, such as perceptions).

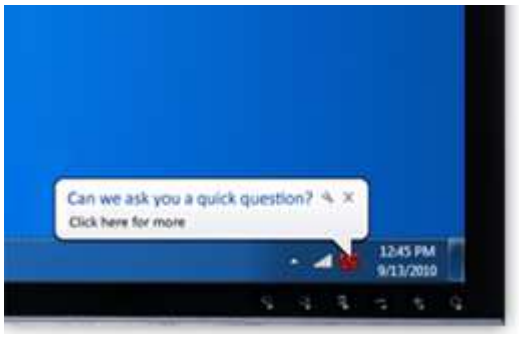


Figure 4 - Dummy screenshot of the event-driven survey module

The first part of the client is already unique in its kind. As far as we know this is the first time that language in actual use will be measured on a world-wide scale, both off line and on line, on a near real-time basis. The combination from the second part with the first part – online survey questions that are automatically triggered by events in the automated language identification agent – is also highly innovative. In short, the user-centric measurements enable very detailed and actual measurement campaigns on the online and offline behavior of users. In DILINET, the behavior specifically refers to the use of language but the instrument could also be used in a more generic manner, to measure many other dimensions as well.

References:

Zhang, J. and M. Van Alstyne (2003). EmailNet: A System for Automatically Mining Social Networks from Organisational Email Communications. North American Association for Computational Social and Organisational Science (NAACSOS), Pittsburgh.

Brennenreadts, R. and R.A. te Velde (2012). Internet as Source of Statistical Data. Final report to the European Commission, DG-Information Society. Utrecht: Dialogic

L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

WP7: Language Indicators in Non-Text Media

Summary:
As in WP5, this work package determines what is available on the web, but searches over audio found in web-based multimedia. As the use of multimedia channels is growing on the web, this work package provides an important dimension, missing from text-based sampling, concerning information available on the web.

Work package WP7 is concerned with developing methods to identify the language in the diverse types of audio data found on websites hosting audio and video documents. As a lot of the information on the web is not in a textual format, these cannot be detected and categorized via text-based methods. A subset of representative data will be selected by the smart sampling techniques developed in WP5, and annotated in WP7 (task 2). An additional challenge that will be addressed is the recognition of highly representative language variants. In coordination with the text-based methods developed in WP8, Task 7.4 will explore techniques to identify the topics and other information in audio data.

Annotating the selected corpus is in and of itself a challenge, as few people are able to accurately tag more than a small number of languages. Therefore an innovative approach will be to use an audio partitioner to chop the audio file into segments, which will then be clustered by speaker and language prior to presentation to humans for annotation. This type of task is well suited to annotation via crowd-sourcing, where segments can be presented to users until a consensus is found. VOCAPIA Research and CNRS will work together to develop this annotation framework. In coordination with Task 7.3, initial systems will be used to reduce the manual annotation load.

Prototype language identification systems will be developed in Task 7.3 and made accessible to the partners via a Web-based service. Performance will be assessed on representative test data selected by the partners, as well as participation in appropriate international campaigns organized outside of the project. Topic or content labels will be annotated in the automatic transcripts for languages which STT systems are available (task 7.4). The automatic transcripts provided via the transcription service will also be used as text input to allow the developed opinion and sentiment analysis methods of WP8 to be applied to audio and audiovisual documents.

WP8: Validity & Analysis

Summary:
This work package provides technical tools for the sampling methods in WP5 to provide clean data (eliminating spam and duplicates), and to provide a first round of annotation (language identification, topic identification, link structure) needed for the sampling methods.

This Work Package interacts with WP5 which develops the sampling techniques for choosing representative web pages, and WP7 which provides automatic transcripts of speech content. In this

WP8, techniques are developed to recognize and eliminate spam, spam farms, duplicate and near duplicate pages. Techniques will also be developed for extracting from the remaining, validated pages, the substantive content of the pages (that is, removing advertisements, header, and footer, and menus).

Task 8.1 involves cleaning the input for the sampling phases of WP5. Web spam is a growing menace to exploiting the unstructured Big Data on the web, since much of the spam involves automatically generated content, designed specifically increase the ranking of pages in search engine results[29]. Search engines engage in a never ending battle to recognize and eliminate these pages, and it is vital to continue this effort as the Web is to be used to detect real patterns and trends. In this WP8, we will continue to develop spam detection algorithms to isolate only content-bearing pages. Spam detection uses different statistical processes between true content and automatically generated content[30]. We can use both the statistical nature of the vocabulary found in the pages, as well as the link structure found between pages, and apply machine learning techniques to these graphs[31]. In DILINET, we will extend current techniques based on graph structure and page content to the different language groups explored in the project. Interacting with the sampling methods of WP5, we will use graph normalization features which include structure and page content, exploring different classification schemes: conditional random fields, support vector machines, and Bayesian classifiers. The main challenge will be to provide tools which are accurate and scale to the large samples to be extracted. We will integrate existing pornography detectors already available within the partner EXALEAD. In addition to eliminating spam, Task 8.1 will perform some low level cleaning of the remaining web pages: removing header and footer information and menus, etc, using standard Readability techniques. It will also perform de-duplication using normalized page signatures (removing number and dates) to provide single versions of cleaned pages during the sampling process.

Language detection, Task 8.2, is an important part of this work package. Current techniques for language detection involve developing statistical models of letter patterns found in words. These patterns can be automatically transliterated into different coding schemes (ISO8859, Windows, KOI8, JIS, UTF8, UTF16, etc), once identified text is gathered for any language. In this work package we will extend language models to all the languages available in Wikipedia (over 200), and from other sources, such as Kevin Scannell's Crubadan project (http://borel.slu.edu/crubadan) which collects samples of rare languages from the web. A major challenge of this work package will be to create a language identifier that is accurate and compact to be exported in the voluntarily downloaded plugin developed in WP6, which must remain light. We believe this will be possible by restricting the language letter sequence model to those patterns that cover a significant portion of possible letter sequences in each language[32].

[29] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, pages 39–47, Chiba, Japan, 2005

[30] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In Proceedings of the seventh workshop on the Web and databases (WebDB), pages 1–6, Paris, France, June 2004

[31] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting spam blogs: A machine learning approach. In Proceedings of the National Conference on Artificial Intelligence(AAAI), Boston, MA, USA, July 2006

[32] T. Gottron and N. Lipka. A comparison of language identification approaches on short, query-style texts. In Advances in Information Retrieval, 32nd European Conference on IR Research (ECIR 2010), pages 611–614, 2010

Figure 5 - Dmoz categories

In addition to flagging spam and identifying languages, this WP, in task 8.3 will also perform gross category recognition (for example, medicine, economics, sports). Category identification involves building up a language model for each category. One commonly used model is the real-world DMOZ concept hierarchy (http://dmoz.org) which uses voluntary editors to create a large directory of classified web pages in more than 300 languages. As with the language identification scheme, the main challenge for DILINET will be to create a small compact categorizer[33] that can be exported in the WP6 plug-in. This will be achieved by restricting the classifiers to a very shallow portion of the DMOZ hierarchy, for example to the bold categories displayed in the image above. A second aspect of Task 8.3 is the multilingual detection of subjective opinions expressed in web pages. Using resources like dictionaries, automatic translation and linked open data two types of opinion mining techniques will be developed by FRAUNHOFER. The first technique will target text sections like paragraphs or whole documents which have been assigned to content categories and estimate their overall subjective orientation (e.g. negative, neutral, positive). The second technique will train models to detect opinions with respect to a selected set of concepts represented by phrases (e.g. "Euro", "Climate Change") by performing a subjective analysis of the text found around these concepts.

As WP 8 gathers URLs and cleaned pages, we will also maintain links into and out of these pages. EXALEAD possesses means of transforming URLS into country codes, combining these country codes and language identification of the pages involved, we will develop in Task 8.4 a map of inter country and intra language linking. UPF will develop tools to detect communities (coherent sub-graphs) in the web graph, using techniques such as Dourisboure (2007)[34] which will allow us to discover language communities in the Web that we can describe by categories produced in Task 8.3.

The results of this work package will be exploited by the deeper content analysis task of WP9: “Data & Knowledge Representation” that will produce the linguistic indicators and language models from this cleaned and categorized data.

[33] Using techniques such as described here: Sujeevan Aseervatham, Anestis Antoniadis, Eric Gaussier, Michel Burlet, Yves Denneulin. A sparse version of the ridge logistic regression for large-scale text categorization. Pattern Recognition Letters, 32(2):101-106, 2011.

[34] Y. Dourisboure, F. Geraci, and M. Pellegrini. Extraction and classification of dense communities in the web. In WWW’07,
pages 461–470, 2007.

WP10: System Development and Integration

Summary:
We develop an infrastructure and platform that coordinates all the technical modules developed in WP5 through WP9 providing a framework for easy integration of the results into runnable prototypes.

A fixed terminology documented in a glossary will be defined to allow unambiguous and clear communication between all partners. We will collect functional and non-functional requirements and document them. The functional requirements describe the DILINET capabilities provided and required by the work packages for solving a functional, application-specific problem. The non-functional requirements refer to system requirements which are not of a functional nature, but contribute decisively to the applicability of the system. They define, e.g. quality requirements, safety and security requirements or performance requirements.

All partners specify the requirements and the outputs for their software modules. Based on these requirements the requirements for the tools, services and overall system are derived. To coordinate this architecture will be developed that allows for easy integration of the different subparts of the system developed in other work packages. To provide open and flexible interfaces which will assure the system's extensibility and acceptability, a detailed analysis of the technological requirements will be performed and possible solutions will be compared to these requirements.

A development process and accompanying tools to support building and integrating the parts will be proposed to enable a smooth workflow for all partners. Furthermore a joint project and compute infrastructure as well as processes and tools required in different work packages will be developed and implemented. The completion of the different project prototypes will be coordinated and supported by providing the infrastructure and common tools.

The result will be three DILINET prototypes which are increasingly comprehensive. The requirements of these prototypes are specified in the System Requirements and Architecture Reports available six months before the prototype releases. Manual inspection of software modules as well automatic software tests and metrics will be employed to generate an up-to-date assessment of functional and non-functional adequacy as well as the software quality of the different software modules provided by the different WPs. Based on these assessments the requirements for the next prototype are updated in a concerted way.

WP12: Results Assessment & Evaluation

Summary:
In this work package, the results and indicators produced by all the methods and modules are crosschecked, evaluated in terms of coherence and verified by comparison with existing proven methods.

DILINET will open various research paths in parallel to obtain data of characterization of the web, particularly language indicators. It is then important to have a single point of analyze and crosschecking of both the process and the produced data in order to help and guide researchers and insure a coherent set of processes and products from the project.

The main concept of WP12 is to evaluate the research from three complementary perspectives: 1) from the user (analyzing the production of data), 2) from the system (analyzing the whole project as a coherent system and checking its software components) and 3) from a methodological soundness approach of the system as a black box combining internal and external views.

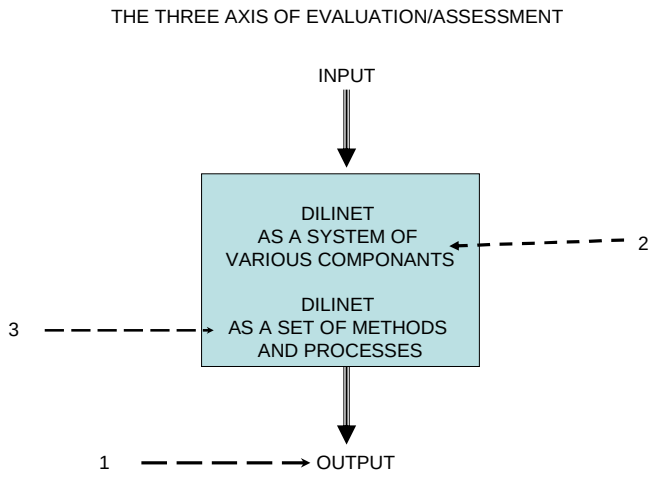


Figure 6 - Evaluation from the user perspective

There are two main lines of verification of the research production: one from comparison with existing proven methods, the other one by analysis of the context and assess with controlled data the likelihood of the results.

As for the first line, the two methods which have produced the most consistent set of results will be used: the one of the Language Observatory Project which has been prone to offer results for minority languages in Asia and Africa and will be applied to Latin America and Europe; and the one of FUNREDES/Union Latine which offered results limited to a subset of 8 languages (Catalan, English, French, German, Italian, Portuguese, Spanish and Rumanian) and can be applied in any defined Internet domain. The LOP method is using script analysis and language recognition applied to a specific crawled Internet space (traditionally ccTLD spaces). The FUNREDES/Union Latine method uses a sample vocabulary of few thousand words which will be searched and counted in web pages in order to determine percentages of language presences compared to English from statistical process. The LOP method will be applied within the project from the crawling approach determined by WP10 and the FUNREDES/Union Latine method which has been in the past used by the means of search engines occurrence counting will be reprogrammed and adapted to be used from crawling. Both methods will provide reference figures with some level of acceptable likelihood and we offer the first alternative for assessment of the data production.

The second alternative for data production will be derived from the analysis of the Internet space which has been object of the result production and determination, from linguistic data about populations, Internet access and any other appropriate verified parameters, of the evaluation of the produced data as well as the providing of elements to help diagnose possible errors. Such approaches have been used successfully in the past with both methods cited above.

Evaluation from the system components perspective

In DILINET a number of processing steps will be implemented which collect the data, perform different analyses and aggregate the results. A large scale data storage is implemented in task 5.1 which stores the web data and panel data and the intermediate results. In addition we will require parallel processing capabilities to execute sampling and analysis within the time limits.
The evaluation from the system components perspective will control if the chosen architecture and the implemented methods really fulfil the requirements of the overall system. A key will be a flexible format for storing data and intermediate results, which easily can be adapted to new requirements and allows

seamless access to the data. Another factor is the modularity of components, which leads relatively autonomous subsystems and allows and easy maintenance and adaption. An important aspect is the overall adequateness of the infrastructure with respect to computing efficiency and storage capacity, as we have to process many hundred TB of data. The evaluation task will monitor the implemented system and give feedback to the partners where necessary.

Evaluation from the methodological perspective

The methodological evaluation assesses both the internal validity and the external of the research results. Note that specific bird-view perspective of this task. Different than the first task, the research results will not be analyzed in detail. Instead, the focus is on the research processes that lead to the results. It is those processes that we will carefully scrutinize. This involves an ex ante audit (are the right processes in place? have sufficient quality safeguards been built in?') and an ex post audit ('have the processes that were [re]defined after the ex ante audit been followed?'). Moreover, there will be a lighter ongoing audit during the entire implementation of the project. These throughput-based audits will be done on a random basis. To ensure strict independency, the work packages were the auditor himself is mainly involved (WP6) will be audited by another partner within to consortium.

With regard to the internal validity, the focus is on the validity of the measurement instruments that is being used. One important part of the ex ante audit to see whether the steps within the processes (measurement instruments can be regarded as a set of activities that are logically connected) have been realistically defined, and with enough details. In the ideal case, every step includes an indicator ('tell-tale') that tells whether the step has been implemented correctly. Obviously, a careful balance has to be struck between the administrative burden to the researchers (and respondents) involved – which should be minimized – and the effectiveness of the audit – which should be maximized. Hence the challenge is to install as little as possible tell-tales but at the most critical places. In order to trace these places, the instrument as a process has to be considered as a whole. This integral assessment of the overall process is the second important element of the ex ante audit.

Most of the methodological evaluation will be based on desk research. Researchers will only be contacted when a signpost surpasses a critical value. This will always be done in close consultation with the WP leader and the overall program manager. The field research involves a fair hearing of the researchers involved. In our experience it could very well be the case that it is the audit framework (which is inherently rather sterile) should be adjusted, and not the implementation (which is done in actual practice).

Whereas internal validity deals with the quality of the measurement instrument, external validity refers to the quality of the samples that have been used. Again applying the overall perspective, this means that we focus particularly on the steps in the processes where the samples are being designed (ex ante audit) and how the sample units (respectively URLs and panelists) are selected and maintained (ongoing audits). The latter is an important issue because due to pragmatic considerations structural biases might creep into the original sample (e.g., a particular category is more difficult to include than others). The ex post audit especially deals with the range of the conclusions and research statements. These should not be overstretched, that is, the range should not extent the actual scope of the underlying samples. This means that in the case of smaller samples than expected, the range of the statements should be narrowed, Likewise, biased samples could be partly adjusted by introducing weightings, but the adjustments should be clearly stated in disclaimers ('we can never be sure that the results are genuine outcomes of the research, or merely constructs due to the particular biases in the samples')

1.3.2.2.2 *Applicative work package overview*

WP4: Societal Issues

Summary:
The principal applicative focus of DILINET's characterization of the web concerns language and language diversity. In this work package, the institutional partners define the societal and linguistic indicators needed by policy makers. Intellectual property and privacy matters are also addressed here.

In this work package, all the side considerations that are required to strengthen the research and insert it into the appropriate societal contexts will be considered. The concept is of a comprehensive approach to enhance the integrity of the products of the developed research and relate them to various relevant societal realms (digital economy, linguistic policies, information society indicators…) in order to obtain impacts from the research beyond the theme of intelligent data management.

The work package will deal with a series of matters which may affect the requirements for the research, the process of the research or the results itself. Those matters could be classified as:

- context setting (linguistic figures),
- protection of all stakeholders interests (intellectual property, legal, ethical, regulatory, security and privacy),
- and correlation (information society indicators).

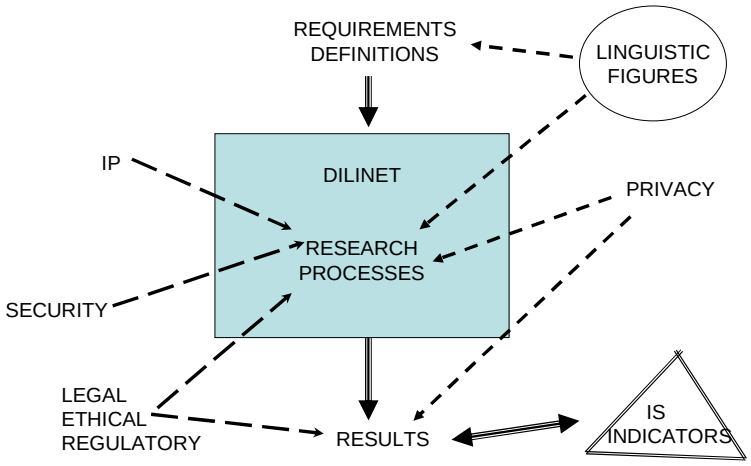


Figure 7 - Focus of WP4 Societal Issues

WP11: Applications

Summary:
The DILINET researches will offer a wide range of possibility to exploit; the selected applications aim to demonstrate these possibilities in a limited but concrete set of situations. We develop then a number of pilot applications exploiting the language indicators and sampling results: local search engines for underrepresented languages, measurement of language diversity in government collections and digital libraries, valuable marketing data for the content industry. From the evaluation of the applications, future impacts of DILINET are forecasted.

The Internet is the fastest medium for individuals to communicate about content, products, services etc. The DILINET will provide insights to get information about language characteristics and important developments to provide individuals the content, products, services etc. they are looking for.

We develop a number of pilot applications exploiting the language indicators and sampling results: local search engines for underrepresented languages, measurement of language diversity in government collections and digital libraries, valuable marketing data for the content industry. From the evaluation of the applications, future impacts of DILINET are forecasted.

WP1: Project Management

This work package provides an efficient administration and co-ordination of all project activities and ensures that the project remains in compliance with its work plan. It liaises with the European Commission to deal with all contractual, financial and administrative matters. It ensures reports and deliverables are submitted on time and to the highest level of quality. It provides the collaborative platform required for maximum synergy between all project teams. It monitors usage of resources and of EC funds.

WP2: Scientific Coordination

This work package ensures the planning, management and monitoring of project-wide research and technological development activities. It carries out regular progress and scientific performance assessments and is in charge of conducting reviews and updates of the Description of Work with a view to maximising project impact. WP2 also coordinates the activities of the Advisory Boards.

WP3: Dissemination & Exploitation

> Summary:
> Dissemination of the language indicators will be performed by organizing international workshops in countries from different continents and language families, developing training tools for exploiting indicators and replicating their collection using the tools developed in the project. There will also be classic dissemination of scientific results through publications. Data produced will be made available on a public web site.

The purpose of this work package is the dissemination and exploitation of the progress and results of the work achieved in the project. A detailed Dissemination and Use Plan as well as an Exploitation Plan, will be compiled to establish the various stages in the process.

Dissemination and Training: The expected results from DILINET will be disseminated in a number of ways, leveraging on the international organisations affiliated to the project. A public DILINET website will contain information on (the progress of) the project, published reports, papers made available to the research community. Planned initiatives include publication in scientific journals, presentations at conferences, networking. Consortium members will also participate in clustering activities, in order to exchange ideas with projects active in the same area.

Exploitation: This includes the planning and start of the technological exploitation of the scientific innovations resulting from the work accomplished in DILINET. Exploitation of the final results of the project will happen via all partners, industrial as well as institutional. A proof of concept will demonstrate the concept of the developed technologies. Plans will be made with regards to marketing and potential implementation and deployment.

Further details on dissemination, exploitation and training are provided in the impact section.

Work package and tasks dependencies

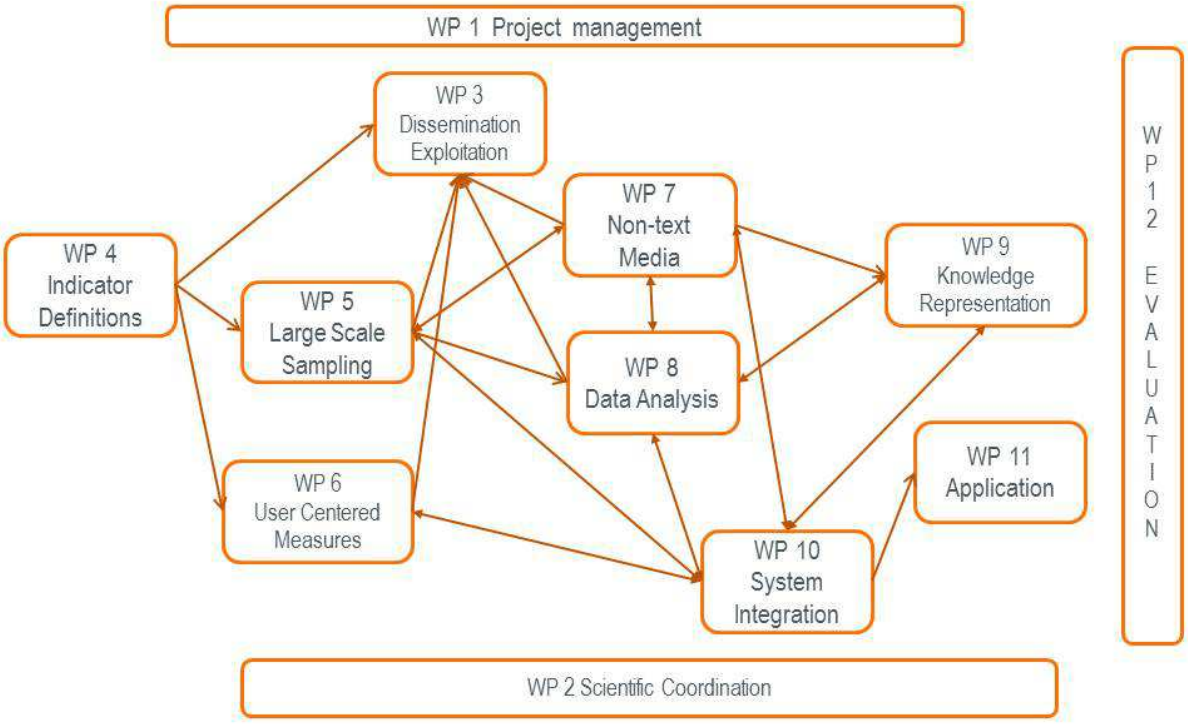


Figure 8 - Work package dependencies

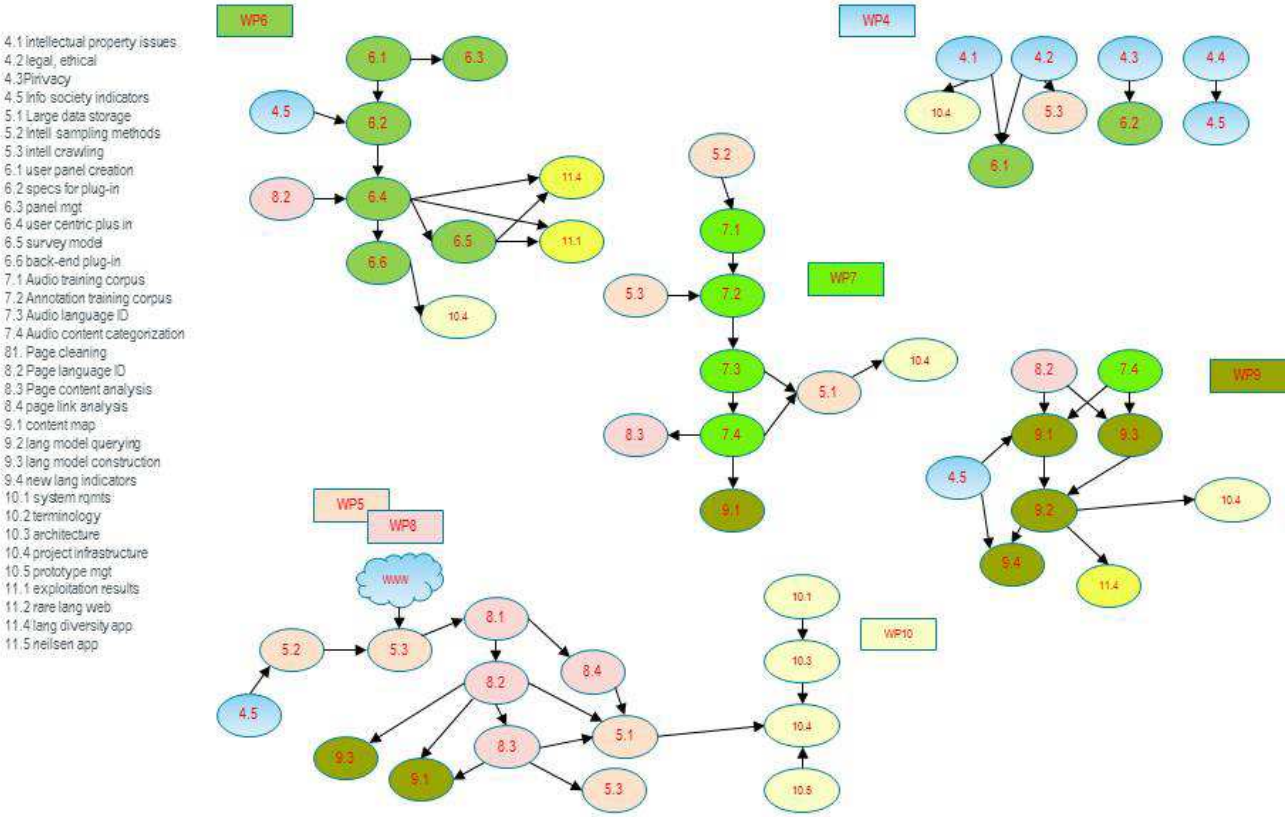


Figure 9 - Task dependencies

Gantt-chart

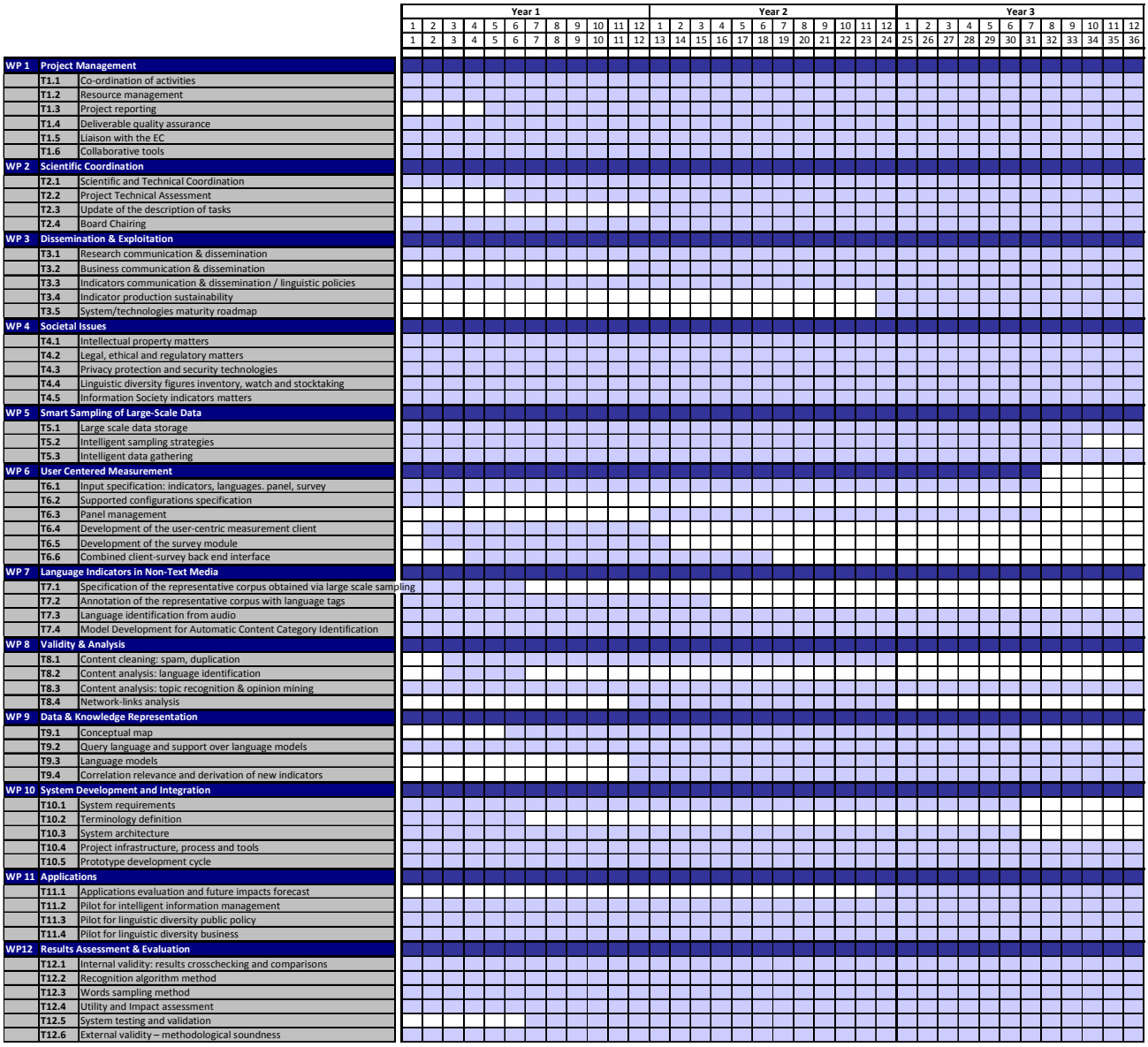


Figure 10 - Gantt chart

B 1.3.3 Work package list and detailed description

Table 1.3a: Work package list

Work package No	Work package title	Type of activity	Lead partic no.	Lead partic. short name	Person-months	Start month	End month
WP 1	Project Management	MGT	1	ERCIM	26	1	36
WP 2	Scientific Coordination	RTD	6	EXALEAD	17	1	36
WP 3	Dissemination & Exploitation	OTH	2	MAAYA	63	1	36
WP 4	Societal Issues	RTD	2	MAAYA	44	1	36
WP 5	Smart Sampling of Large-Scale Data	RTD	3	CNR	109	1	36
WP 6	User Centered Measurement	RTD	4	DIALOGIC	54	1	36
WP 7	Language Indicators in Non-Text Media	RTD	5	CNRS	101	1	36
WP 8	Validity & Analysis	RTD	6	EXALEAD	95	1	36
WP 9	Data & Knowledge Representation	RTD	9	CWI	52	1	36
WP 10	System Development and Integration	RTD	8	FRAUNHOFER	103.4	1	36
WP 11	Applications	RTD	13	NIELSEN	73	1	36
WP12	Results Assessment & Evaluation	RTD	10	FUNREDES	38	1	36
	TOTAL				775.4		

Figure 11 - Work package detailed list

Table 1.3b: Deliverables List

Del. no.	Deliverable name	WP no.	Nature	Dissemi-nation level	Delivery date
D1.1	Collaborative working environment documentation	1	R	PU	M6
D1.2	First bi-annual activity report	1	R	PU	M7
D1.3	Periodic report Year 1	1	R	PU	M12
D1.4	Second bi-annual activity report	1	R	PU	M19
D1.5	Periodic report Year 2	1	R	PU	M24
D1.6	Third bi-annual activity report	1	R	PU	M31
D1.7	Periodic report Year 3	1	R	PU	M36
D1.8	Final project reports	1	R	PU	M36
D2.1	Scientific sections of the Periodic Report Year 1	2	R	PU	M12
D2.2	Scientific sections of the Periodic Report Year 2	2	R	PU	M24
D2.3	Scientific sections of the Periodic Report Year 3	2	R	PU	M36
D2.4	Scientific sections of the Final Project Reports	2	R	PU	M36
D3.1	Web site Version 1	3	O	PU	M6
D3.2	Web site version 2	3	O	PU	M12
D3.3	Web site version 3	3	O	PU	M18
D3.4	Web site version 4	3	O	PU	M24
D3.5	Design of training package	3	O	PU	M24
D3.6	Web site version 5	3	O	PU	M30
D3.7	ITU conference	3	O	PU	M30
D3.8	Special events at the international conferences	3	O	PU	M30
D3.9	First series of publications	3	R	PU	M30
D3.10	National capacity building workshops	3	O	PU	M30
D3.11	Strategy document for research roadmap	3	R	PU	M36
D3.12	Strategy document for indicator sustainability	3	R	PU	M36
D3.13	Second series of publications	3	R	PU	M36
D3.14	Web final version	3	O	PU	M36
D4.1	Document of understanding for IP matters	4	R	PU	M6
D4.2	Document of understanding for legal matters	4	R	PU	M6
D4.3	Document of understanding for ethical matters	4	R	PU	M6
D4.4	Linguistic figures version 1 (plus inputs for indicators specs)	4	R	PU	M6
D4.5	Document of recommendations for security matters in WP6	4	R	PU	M12
D4.6	Linguistic figures version 2	4	R	PU	M12
D4.7	Stress test reports of security for WP6	4	R	PU	M24
D4.8	Linguistic figures version 3				M24
D4.9	Strategy document for information society indicators	4	R	PU	M30
D4.10	Strategy document “languages on the Internet” for standard bodies	4	R	PU	M30
D4.11	Linguistic figures final version	4	R	PU	M36
D5.1	Requirements and specifications for the Large scale data storage.	5	R	PU	M6
D5.2	Requirements and specifications for the Intelligent sampling strategies.	5	R	PU	M6
D5.3	Requirements and specifications of the Intelligent data	5	R	PU	M18

	gathering system				
D5.4	Interactions of the Intelligent sampling with the data gathering systems and how data are efficiently stored in the large scale data storage.	5	R	PU	M24
D5.5	Implementation of a prototype of the large scale data storage module.	5	R	PU	M36
D6.1	Report on scope of the project in terms of number and types of operating systems supported	6	R	PU	M3
D6.2	First stable version of protocol	6	R	PU	M3
D6.3	Mailing list of panel (initial set of respondents)	6	R	RE	M12
D6.4	Final stable version of protocol	6	R	PU	M22
D6.5	User-centric measurement client, including survey module and back end interface	6	P	RE	M36
D6.6	Final report based on the analysis of the user-centric measurements	6	R	PU	M36
D7.1	Corpus Specification	7	R	RE	M3
D7.2	Report on Annotated Corpus	7	R	RE	M12
D7.3	Report on LID systems v1	7	R	RE	M6
D7.4	Report on Topic Detection in audio v1	7	R	RE	M14
D7.5	Report on LID systems v2	7	R	PU	M36
D7.6	Report on Topic Detection in audio v2	7	R	PU	M36
D8.1	First requirements for category analysis	8	R	PU	M6
D8.2	First prototype for category analysis	8	P	RE	M12
D8.3	Updated requirements for category and opinion analysis ...	8	R	PU	M18
D8.4	Second prototype for content and opinion analysis	8	P	RE	M24
D8.5	Updated requirements for content and opinion analysis	8	R	PU	M30
D8.6	Third prototype for content and opinion analysis	8	P	RE	M36
D9.1	Query Language Specification (T9.2)	9	R	PU	M9
D9.2	Tool for Conceptual Maps V1 (T9.1)	9	P	RE	M12
D9.3	User Study – Conceptual Maps (T9.4)	9	R	PU	M18
D9.4	Language Models Publicly Available (T9.3)	9	R	PU	M30
D9.5	Interactive Tool (T9.2)	9	P	RE	M24
D9.6	User Study – Conceptual Maps with topics/opinions (T9.2 & T9.4)	9	R	PU	M30
D9.7	Tool for Conceptual Maps V2 (T9.1, using T9.2)	9	P	RE	M36
D10.1	Terminology Definition Report	10	R	PU	M2
D10.2	System Requirements Report	10	R	PU	M4
D10.3	First System Architecture	10	R	PU	M6
D10.4	First Prototype	10	P	RE	M12
D10.5	Updated System Requirements and Architecture Report	10	R	PU	M18
D10.6	Second Prototype	10	P	RE	M24
D10.7	Final System Requirements and Architecture Report	10	R	PU	M30
D10.8	Third Prototype	10	P	RE	M36
D11.1	Applications evaluation framework	11	R	PU	M24
D11.2	Intermediary Report from intelligent information management pilot	11	R	PU	M30
D11.3	Intermediary Report from linguistic diversity public policy pilot	11	R	PU	M30
D11.4	Intermediary Report from for linguistic diversity business pilot	11	R	PU	M30

D11.5	Final Report from intelligent information management pilot	11	R	PU	M36
D11.6	Final Report from linguistic diversity public policy pilot	11	R	PU	M36
D11.7	Final Report from for linguistic diversity business pilot	11	R	PU	M36
D11.8	Impacts forecast Report	11	R	PU	M36
D12.1	Software development to adapt FUNREDES/UL method to crawling	12	P	RE	M12
D12.2	Reference measurement made with FUNREDES/UL method	12	R	PU	M18
D12.3	Reference measurement made with LOP method	12	R	PU	M18
D12.4	Result analysis and assessment first report	12	R	PU	M18
D12.5	System validation first report	12	R	PU	M18
D12.6	Methodology assessment first report	12	R	PU	M18
D12.7	Result analysis and assessment second report	12	R	PU	M24
D12.8	System validation second report	12	R	PU	M24
D12.9	Methodology assessment second report	12	R	PU	M24

Table 1.3c: List of milestones

Milestone number	Milestone name	WP(s) involved	Expected date	Means of verification
M3.1	Consolidated version of web site	3	M12	Online check.
M3.2	Consolidated plan for all dissemination and exploitation activities	3	M18	Partner meeting
M3.3	Review of plan for all dissemination and exploitation activities	3	M24	Partner meeting
M3.4	Evaluation of all dissemination and exploitation activities	3	M36	Report
M4.1	Partner's meeting for consensus on legal, ethical, privacy and security matters	4	M12	Meeting evaluation
M4.2	Partner's meeting for finalization on specification for indicators	4	M12	Meeting evaluation
M4.3	Partner's meeting for strategy decisions for languages matters and information society	4	M36	Meeting evaluation
M5.1	Specifications of the Large scale data storage architecture	5	M9	Specifications available
M5.2	First implementation of the Large scale data storage including intelligent sampling and data gathering techniques	5	M24	Large scale data storage V1 available
M5.3	Final implementation of the Large scale data storage.	5	M36	Final large scale data storage available
M6.1	Baseline protocol ready	6.1	M5	Protocol has been thoroughly tested internally and tested successfully on an external panel
M6.2	Baseline panel ready	6.1	M12	A panel has been composed that

				meets the critical statistical size in each country involved
M6.3	Beta version 1.0 of client ready	6.4	M12	First version of the client available, technically tested (bug free)
M6.4	Beta version 1.1 of client ready	6.5	M13	Second version of the client available, with survey module integrated, technically tested (bug free)
M6.5	Panel active for 9 months (halfway full duration)	6.3	M22	Client installed at devices from 10,000 users in 10 countries, with a consistent 90% of active users
M6.6	Debriefing of panel (end of duration, after 18 months)	6.3	M31	All 10,000 users informed about the completion of the measurement.
M7.1	Baseline web-based LID	7	M6	service available
M7.2	Baseline topic detection web-based	7	M12	service available
M7.3	Extended web-based LID	7	M18	service available
M7.4	Final web-based LID	7	M34	service available
M7.5	Final topic detection web-based	7	M34	service available
M8.1	First prototype of category analysis	8	M12	software evaluation and test data
M8.2	Second prototype of category and opinion analysis	8	M24	software evaluation and test data
M8.3	Second prototype of category and opinion analysis	8	M36	software evaluation and test data
M9.1	Pilot of first user study completed	9	M15	results of pilot study available
M9.2	Language Models Prepared for Internal Use	9	M24	A demo application based on statistical language modelling, e.g. word breaking, uses the DILINET models
M9.3	Pilot of second user study completed	9	M27	results of pilot study available
M10.1	First Prototype	10	M12	Automatic software tests and metrics as well as manual inspection according to requirements.
M10.2	Second Prototype	10	M24	Automatic software tests and metrics as well as manual inspection according to requirements.
M10.3	Third prototype	10	M36	Automatic software tests and metrics as well as manual inspection according to requirements.
M11.1	Consensuated Evaluation Framework and launch of Applications	11	24	Partner meeting
M11.2	Intermediary evaluation and reporting	11	30	Advisory Boards evaluation of reports
M11.3	Final reporting	11	36	Final meeting with presence of Advisory Boards
M12.1	FUNREDES/UL method adaptation	12	M12	Software test in sample data.
M12.2	Reference measurements obtained	12	M18	Crosschecking between T12.2 and T12.3
M12.3	Assessment first reports	12	M18	Partner meeting
M12.4	Assessment second reports	12	M24	Partner meeting

Table 1.3d: Work package description

Work package number	**1**		**Start - End:**		M1 – M36				
Work package title	**Project Management**								
Activity type[35]	MGT								
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	**ERCIM**	MAAYA	CNR	DIALOGIC	CNRS	EXALEAD	UPF	FHG	CWI	FUNREDES
Person-months per participant	**22**	1								3
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant										

Objectives

The objective of this work package is to ensure an efficient administration and co-ordination of all project activities, focused to the objectives of the project.

Description of work

Task 1.1 – Co-ordination of activities (M1-M36, 8 PM)

Task Leader: ERCIM

Participants:

This task is to plan and monitor the project activities, in close collaboration with the Scientific Coordinator, ensuring an effective coordination, detecting early possible deviations, and appropriately addressing these. It includes the organisation and participation in (physical or virtual) project meetings. Specific measurable indicators will be defined in order to monitor overall project progresses with respect to the global project objectives. This task includes monitoring of non-technical risks and taking the necessary actions for risk mitigation.

In order to run the project in a professional manner, the procedures that will be put in place will be described in "Management Notes" that will be available in the Collaborative Web Environment.

This task consists also of preparing the Consortium to official Project review meetings, ensuring timely provision of information, good quality and professional presentations and rehearsal sessions.

The coordinator will chair the General Assembly.

Task 1.2 – Resource management (M1-M36, 3 PM)

Task Leader: ERCIM

Participants:

This task is to manage the project resources and specifically the project budget. It consists of administrating the Community financial contribution, monitoring the use of resources by each partner,

[35] Please indicate one activity per work package:
RTD = Research and technological development; DEM = Demonstration; MGT = Management of the consortium; OTHER = Other specific activities if applicable to this call, including any activities to prepare for the dissemination and/or exploitation of project results and coordination activities.

and validating transfer of budget between activities and beneficiaries.

Task 1.3 – Project reporting (M5-M36, 4 PM)

Task Leader: ERCIM

Participants:

This task is to produce the set of deliverables to the European Commission on a regular basis. The task consists of, on one end to gather the contribution by all partners, and on the other end to produce the periodic and final project reports.

The periodic reports shall comprise:

- An overview, including a publishable summary of the progress of work towards the objectives of the project, including achievements, milestones and deliverables). If there were deviations from the original work plan, this report includes their description and motivation.
- An explanation of the use of the resources.
- A Financial Statement from each partner together with a summary financial report consolidating the claimed Community contribution of all the project partners in an aggregate form.

This final report shall comprise:

- A final publishable summary report covering results, conclusions and socio-economic impact of the project.
- A report covering the wider societal implications of the project, including gender equality actions, ethical issues, efforts to involve other actors and to spread awareness, as well as the plan for the use and dissemination of foreground.

All reports and deliverables will be published in English.

Task 1.4 – Deliverable quality assurance (M1-M36, 3 PM)

Task Leader: MAAYA

Participants: ERCIM

In collaboration with he Project Co-ordinator, the task leader will organise and drive a quality assurance process for all project deliverables. The process will be as follows:

- Intended table of content of the deliverable is available 1 month after the corresponding task kick-off date.
- Draft deliverable versions are available four weeks before planned delivery date.
- Appointment of two internal reviewers from different organisations than the deliverable author(s).
- Reviewers carry out their reviews and produce a written (email) report within two weeks of reception of the draft deliverable.
- Author(s) take into account the reviewers comments and produce the final version of the deliverable.

This procedure will be documented as a “Management Note”. In all project management meetings, MENON will report on the quality assurance process to make sure that the group improves its performance.

Task 1.5 – Liaison with the EC (M1-M36, 3 PM)

Task Leader: ERCIM

Participants:

This task is to liaise project information and progresses with the Project Officer at the European Commission. A communication channel will be continuously kept open, via electronic mail, telephone, conference calls or physical meetings. It consists also of supporting the Project Officer in organising the official project reviews and make sure the feedback from the reviewers is appropriately addressed by the project.

Specifically, any significant change in the planned activities (change in the Description of Work, transfer of budget between activities and beneficiaries) will be formally requested to the Project Officer.

Task 1.6 – Collaborative tools (M1-M36, 5 PM)
Task Leader: ERCIM
Participants: FUNREDES
In order to support an efficient collaboration between the project partners, we will set up and manage an appropriate project IT infrastructure, that will consist of 5 components:

- Mailing lists,
- Wikis,
- Collaborative web environment (BSCW or equivalent),
- Source code Revision Management,
- Issue Tracker.

Deliverables (brief description) and month of delivery					
Number [36]	Description	WP	Nature[37]	Diss level [38]	Month Due
D1.1	Collaborative working environment documentation	1	R	PU	M6
D1.2	First bi-annual activity report	1	R	PU	M7
D1.3	Periodic report Year 1	1	R	PU	M12
D1.4	Second bi-annual activity report	1	R	PU	M19
D1.5	Periodic report Year 2	1	R	PU	M24
D1.6	Third bi-annual activity report	1	R	PU	M31
D1.7	Periodic report Year 3	1	R	PU	M36
D1.8	Final project reports	1	R	PU	M36

Milestones				
Number	Short description	WP	Month	Means of verification[39]
N/A				

Note: footnotes commenting the WP description table headings have been left visible on WP1 only.

[36] Deliverable numbers in order of delivery dates.
[37] Please indicate the nature of the deliverable using one of the following codes:
R = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other
[38] Please indicate the dissemination level using one of the following codes:
PU = Public
PP = Restricted to other programme participants (including the Commission Services).
RE = Restricted to a group specified by the consortium (including the Commission Services).
CO = Confidential, only for members of the consortium (including the Commission Services).
[39] Show how you will confirm that the milestone has been attained. Refer to indicators if appropriate. For example: a laboratory prototype completed and running flawlessly; software released and validated by a user group; field survey complete and data quality validated.

Work package number	**2**		**Start - End:**		M1 – M36				
Work package title	**Scientific Coordination**								
Activity type	RTD								
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	DIALOGIC	CNRS	**EXALEAD**	UPF	FHG	CWI	FUNREDES
Person-months per participant						13	4			
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant										

Objectives

This Work Package is to scientifically animate the project and make sure synergies between RTD activities and tasks are developed. It aims also at ensuring the highest scientific quality of the research activities and to maximise the chances of delivering the highest impact as possible.

Description of work

Task 2.1 Scientific and Technical Coordination **(M1-M36, 9 PM)**

Task Leader: EXALEAD

Participants: UPF

This task is to carry out the **planning**, **management** and **monitoring** of project-wide research and technological development activities, including the coordination of scientific and technical work between work packages. Special attention will be given to monitoring technical risks and taking the necessary actions for risk mitigation.

Day-to-day technical management of individual work packages, including the coordination of work between other work packages and tasks of a work package is the responsibility of the WP leaders.

Task 2.2 Project Technical Assessment **(M6-M36, 4 PM)**

Task Leader: EXALEAD

Participants: UPF

This task consists of assessing the project progresses along two different axes:

- Assessment of project scientific and technological progresses in comparison with the overall project objectives;
- Assessment of the scientific and technical contribution of the individual project partners.

The objective of this assessment task is to detect as early as possible deficiencies (including underperforming partners) in the project execution or organisation and define corrective actions, in close cooperation with the Project Coordinator.

Task 2.3 Update of the description of tasks **(M12-M36, 2 PM)**

Task Leader: EXALEAD

Participants: UPF

At the end of the first and second year of the project, an audit of the project achievements, objectives

and expected impact will be performed and on basis of the conclusion, the Description of Work (DoW) will be updated to maximise the potential impact of the project.

Task 2.4 Board Chairing **(M1-M36, 2 PM)**
Task Leader: EXALEAD
Participants: UPF
This task consists of planning, organising, preparing (documents, agenda, logistics, etc.) and chairing the management bodies of the project according to the rules defined in the Grant Agreement and in the Consortium Agreement; namely, the Project Executive Board and the Advisory Board. It is anticipated that:

- The scientific coordinator will chair the Project Executive Board
- Yahoo! Research will chair the external Scientific Advisory Board
- ITU will chair the external Societal Advisory Board

Deliverables (brief description) and month of delivery

Number	Description	WP	Nature	Diss level	Month Due
D2.1	Scientific sections of the Periodic Report Year 1	2	R	PU	M12
D2.2	Scientific sections of the Periodic Report Year 2	2	R	PU	M24
D2.3	Scientific sections of the Periodic Report Year 3	2	R	PU	M36
D2.4	Scientific sections of the Final Project Reports	2	R	PU	M36

Milestones

Number	Short description	WP	Month	Means of verification
N/A				

Work package number	**3**		**Start - End:**		M1 – M36				
Work package title	**DISSEMINATION & EXPLOITATION**									
Activity type	OTHER									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	**MAAYA**	CNR	DIALOGIC	CNRS	EXALEAD	UPF	FHG	CWI	FUNREDES
Person-months per participant	5	15	1	2	4	2	4	1	1	8
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant	3	15	2							

Objectives

This work-package is conceived as the main guarantee of the return on the research investment into visible and exploitable products with managed impacts and the corresponding dissemination and exploitation of both the research results and the impacts, among all relevant stakeholder categories. The dissemination will be realized both directly in a comprehensive web site reflecting the progress and outcomes of the research and indirectly through appropriate publications targeting the research and the business communities as well as public at large. The policy side will be specially developed in coordination with International Organisations involved with the production of policy and strategy papers, the organisation of workshops for capacity building, international conferences and production of training materials.

Tangible outcomes and measures of progress and success

The tangible outcomes of this work packages encompasses:
- a web site which shall become a reference on the processed subjects and which measure of progress will be the incoming traffic and the number of links to.
- a set of publications covering the results of the research realized, the impact of the project in terms of e-business, the impact of the project in terms of public policies for linguistic diversity and finally in terms of information society indicators; the success of the publication will be measured by the number of citations;
- a set of position and strategic papers related to the themes of the project which value will be assessed by the External Advisory Board;
- a set of training materials under responsibility of UNESCO for the management of linguistic diversity in the digital world which value will be assessed by the External Advisory Board and which will be evaluated by the users;
- two events at the international level organized respectively by UNESCO and ITU whose indicator of success will be related to the quality and quantity of attendees and the evaluation feed-back;
  - a series of national workshops organized by UNESCO for selected national organisations which will be evaluated by the attendees;
  - a strategy for the sustainability of the indicators production which will be assessed by the

External Advisory Board;
- a Research Roadmap for Internet linguistic diversity measurement which will be assessed by the External Advisory Board.

Description of work

Task 3.1 Research communication & dissemination **(M1-M36, 13.5 PM)**
Task Leader: UPF
Participants: CWI, FRAUNHOFER, CNR, CNRS, DIALOGIC, ERCI, VOCAPIA

Creation and maintenance of the research section of the web site (note that the overall design and hosting will be insured by ERCIM). Coordination of a set of scientific publications reporting to the research community on the research outcomes of the project.
The technical partners and user partners are strongly committed to dissemination of results through well-recognized scientific channels. All have excellent records of publication at major scientific venues in their respective fields, conferences (ISCA Interspeech and workshops, IEEE ICASSP, ASRU, LREC, HLT, etc) and journals (Speech Communication, Computer Speech and Language, Natural Language Engineering, etc). The interdisciplinary nature of the project will provide an important opportunity for the partners to inform experts in fields beyond their own about the impact of the work and the collaboration.

Task 3.2 Business communication & dissemination **(M12-M36, 9.5 PM)**
Task Leader: EXALEAD
Participants: NIELSEN, DIALOGIC, VOCAPIA, ERCIM, CWI

Creation and maintenance of the business section of the web site. Coordination of a set of white paper publications reporting to the business community on the business outcomes of the project. Presentation of the business oriented applications in appropriate venues such as Online Information, WWW Conference, Content Management, Semantic Technologies, IFRA, Social Media World Forum, etc.

Task 3.3 Indicators communication & dissemination / linguistic policies **(M1-M36, 18 PM)**
Task Leader: MAAYA
Participants: UNESCO

Creation and maintenance of the "products" section of the web site which will include practical tools to serve as unique clearinghouse in the subject of linguistic diversity in the digital world. Coordination of a set of publications on the outcomes of the project as far as linguistic diversity in the digital world is concerned, with specific attention devoted to reaching non-expert stakeholders groups. Production of policy and strategy papers as well as technical guidelines for use by EU and involved international organisations. Organisation of workshops and consultations for selected national governmental organisations on the subject of linguistic diversity in digital world. Capacity building of selected national governmental organisations responsible both for the national statistics and information policies through seminars and through provision of online resources. Production of training materials for higher educational institutions in Open Educational Resources (OER) format, in compliance with the UNESCO recommendations in the field. Organisation of two special events at international level to be organized to present the outcomes of the project and share experience with other experts working in this field, one under the responsibility of UNESCO, on linguistic diversity, one under the responsibility of ITU, on information society indicators. Both events will be followed by publications under respective responsibility of UNESCO and ITU. Whenever schedules allow, the ITU conference will be linked and/or

integrated to the celebration of the WSIS+10 event in 2014.

Task 3.4 Indicator production sustainability **(M24-M36, 16 PM)**
Task Leader: MAAYA
Participants: UNESCO, FUNREDES

This task addresses the issue of the sustainability of the production of indicators with the main institutional partners. It will investigate the possibility of enlarging the DILINET community involving relevant stakeholders and specifically users groups and to set up collaboration activities with similar and complementary projects, clusters and programs. It will guarantee a correct and transparent strategic governance of the network and prepare/set up the relevant coordination mechanisms, along a long-term strategic view over the next 5-10 years that will be collaboratively produced at the beginning of the project. It will conduct a foresight exercise on the future research strands in the project field and establish a roadmap for future research on the issue of linguistic diversity on the digital world and finally provide policy and research recommendations, aiming at making sure that the research results of the project are taken up by the appropriate research communities and have an impact on policy in the field beyond the project lifetime.

Task 3.5 System/technologies maturity roadmap **(M24-M36, 6 PM)**
Task Leader: UPF
Participants: MAAYA, FUNREDES

A foresight exercise will be developed to consider future directions for R&D to measure linguistic diversity in the digital world. The topics to be investigated will include: relationship between research, policy and practice; identifying common challenges and key issues facing programmes; ideas for future R&D co-operation; new opportunities and gaps in R&D. Based on the results of this foresight exercise and other relevant work-packages a Research Roadmap for internet linguistic diversity measurement will be produced.

Deliverables (brief description) and month of delivery

Number	Description	WP	Nature	Diss level	Month Due
D3.1	Web site Version 1	3	O	PU	M6
D3.2	Web site version 2	3	O	PU	M12
D3.3	Web site version 3	3	O	PU	M18
D3.4	Web site version 4	3	O	PU	M24
D3.5	Design of training package	3	O	PU	M24
D3.6	Web site version 5	3	O	PU	M30
D3.7	ITU conference	3	O	PU	M30
D3.8	Special events at the international conference	3	O	PU	M30
D3.9	First series of publications	3	R	PU	M30
D3.10	National capacity building workshops	3	O	PU	M30
D3.11	Strategy document for research roadmap	3	R	PU	M36
D3.12	Strategy document for indicator sustainability	3	R	PU	M36
D3.13	Second series of publications	3	R	PU	M36
D3.14	Web final version	3	O	PU	M36

Milestones				
Number	Short description	WP	Month	Means of verification
M3.1	Consolidated version of web site	3	M12	Online check.
M3.2	Consolidated plan for all dissemination and exploitation activities	3	M18	Partner meeting
M3.3	Review of plan for all dissemination and exploitation activities	3	M24	Partner meeting
M3.4	Evaluation of all dissemination and exploitation activities	3	M36	Report

Work package number	**4**		**Start - End:**		M1 – M36				
Work package title	**SOCIETAL ISSUES**								
Activity type	RTD								
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	**MAAYA**	CNR	DIALOGIC	CNRS	EXALEAD	UPF	FHG	CWI	FUNREDES
Person-months per participant		25		1						12
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant		6								

Objectives

This work package gathers the focuses on all the side considerations that are required to guarantee the quality and validity of the research and their insertion onto the appropriate societal contexts. The side considerations are concerned with ethical and legal matters as well as those related to Intellectual Property, security, and privacy. As for the appropriate contexts that are considered one is the linguistic diversity indicators in the digital world which requires to be contrasted with general data about linguistic diversity, and the other ones is the information society complex of indicators of which it is part and which paradigm shall be positively impacted by the progress made in the project. This comprehensive approach is conceived as a contribution to guarantee the integrity of the products of the developed research and relate them to real life practical and useful environments with projection for impact beyond the initial subject of intelligent data management.

Tangible outcomes and measures of progress and success

The tangible outcomes includes:

- a consortium agreement for intellectual property matters which will be evaluated by the External Advisory Board;
- a guidelines document on legal ad ethical matters which will be evaluated by the External Advisory Board;
- a reference document for the securiy and privacy matters which will serve as input for WP6 will be evaluated by the External Advisory Board;
- a report from Kyos of the result of stress testing of security and privacy for the software provided in WP6 will be evaluated by the External Advisory Board;
- a compilation of data and figures about linguistic diversity which will be actualized each 6 months and organized in a section of the web site will be evaluated by the External Advisory Board;
- a report of the interaction of the project with the WSIS process and especially the task group in measuring WSIS targets will be evaluated by the External Advisory Board.

Description of work

Task 4.1 Intellectual property matters **(M1-M36, 3 PM)**

Task Leader: MAAYA
Participants:

This work-package will address all kinds of intellectual property and know-how issues connected to the project that will be generated during the project and as a result of the same. In the line of maximizing the public domain decision when applicable and sharing the rights within the consortium. Intellectual property further comprises participants' pre-existing intellectual property and know-how owned by the participants before the start of the project, and also intellectual property and know-how created outside of the project, during its duration, and which is connected with the project. Proper intellectual property protection will be considered, also under the perspective of possible copyright protection, patentability and any other king of intellectual property protection, in relation to software and more generally any kind of know-how that will be produced in the course of the project, as a result of the same or that will be comprised in the outputs of the project. To this regard, know-how and any kind of intellectual property right developed in relation to a specific stage of the project or as an output of the same will be protected through appropriate procedures and agreements established among the members of the Consortium or will be defined as creative common when appropriate. Management and protection of knowledge and intellectual property will be eased, within the project, by the tight interaction with the legal consulting party. From the early stage until the end of the project time, the legal department of the partners will provide continuous assistance as to management of the knowledge produced and protection of intellectual property rights in any way arising or connected with the project. The partners will determine the appropriate knowledge management procedures and rules within the Consortium at the various stages of work and thereafter, especially for what concerns the innovation aspects of this project. The Consortium Agreement will devote specific and significant attention to the issue of intellectual property rights management. In the Consortium Agreement they will be defined and specified procedures and rules for a proper handling, ownership, managing, protection and granting of the knowledge and of any relevant intellectual property rights, in any way produced and of any kind, with regards to both internal usage for scopes within the project frame, usage outside the project during the project time frame and usage after project completion.

Task 4.2 Legal, ethical and regulatory matters **(M1-M36, 6 PM)**
Task Leader: MAAYA
Participants: UNESCO

This work-package will address any legal, regulatory or ethical issue raised by the research conducted in the project as well as by the products. Particular attention will be given to the User Centered Measurement Programs: the panels will be explicitly Informed (through End User License Agreement/Terms of Use) about legal implications and ethical measures taken to ensure the protection of personal data and the statistical usage of the collected data within the framework of the project. This work package will coordinate with all the partners' guidelines on ethical use of information for the project describing how data will be collected, and why and how it will be used. The fundamental principles outlined in various EU and UN international normative documents as human dignity, integrity of the person, the right to privacy, etc. will be analyzed in order to be fully respected and promoted and feed task 4.3. As a fundamental complement of the technical requirement specification, a thorough analysis and profiling of the legal and regulatory framework will be provided, including both i) privacy and data security requirements set forth at a European level, and ii) privacy laws in selected EU and other countries. These will not be static profiles but will also take into account the position held by local data protection Commissioners and actual practice of the selected countries (in some territories, actual practice differs from written law). A similar review of relevant law enforcement legislation will be undertaken to identify additional obligations and uniform technical models and standardised best practices. Rather than just provide an overview of all the rules in all the data protection laws and secondary rules and regulations, aim of this activity is to describe in a comparative and analytical way the laws in the selected Member States, in order to provide technical requirements

which will ensure that the developed product will comply with the relevant rules of the EU regulatory regime. This will assist to get insights and solutions to achieve a high level of integration between technical concepts and European laws and regulatory provisions.

Task 4.3 Privacy protection and security technologies **(M1-M36, 6 PM)**
Task Leader: MAAYA
Participants: DIALOGIC

The design of the User Centered Measurement programs will be checked carefully against the capture of neither unwanted information nor identifiable data (such as IP number). The risk to see the program be taken control by specific viruses will be scrutinized and the appropriate protection to avoid that risk be put in place. The architecture of the user centered measurements implies a client which is locally installed on one or more devices of the panellist, a central server where all the aggregated information is being stored, and a CRM-system that manages all the interactions with the panellists. The latter includes the sending of targeted survey questions to specific panellists via the client. There is thus two-way data exchange between the local client and the central server and CRM-system. Task 4.3 will audit and resolve the privacy and security risks that are involved in the data exchange to and from the client. Basically there are four types of risks: the identity of the panellist can be deduced from the data that is being sent from the client, data can be intercepted during the exchange, unwanted data can be sent to the client (including malevolent scripts that can be used to take over the client) and the security hole in the server can be exploited which would allow a hacker to compromise the system. This WP will suggest appropriate protection measures to be implemented, respectively for the anonymisation of data (e.g., matching panellists on a hash table, not on the IP address), for the encryption of data, and for the fencing of the client and the server.

Task 4.4 Linguistic diversity figures inventory, watch and stocktaking **(M1-M36, 19 PM)**
Task Leader: MAAYA
Participants: FUNREDES

The activity consists in gathering and compiling all required context data about linguistic matters produced outside the project (such as demographics and country policy assessment) to guide the research required outputs. This will serve as a pool services to other work-packages requirements on linguistic matters and will provide the corresponding content of the web site to transform it into a clearinghouse on the subject of linguistic diversity on the digital world. This task will also input WP12 with appropriate data and chronological series to allow assessment of the results produced by the research. A systematic study and documentation of the state of the art in the subject of linguistic diversity in the digital world will be conducted, considering both aspects relevant to public policies and business. This implies organizing figures for languages in the real world to serve as reference data. The product of this work-package will input the objectives of the research activities with a selection of required indicators, while a permanent watch will be organized and fed back in the corresponding project web sections.

Task 4.5 : Information Society indicators matters **(M1-M36, 10 PM)**
Task Leader: FUNREDES
Participants: UNESCO, MAAYA

The production of linguistic diversity indicators will be set in the more broader context of the production of Information Society indicators which the project will interface in this work package through institutional partners (ITU, MAAYA, UNESCO). It will in particular address, within the follow-up of the World Summit of Information Society (WSIS), the WSIS target 9, which is to “encourage the development of content and put in place technical conditions in order to facilitate the presence and use of all world languages on the Internet”, and WSIS action line C8 (cultural and linguistic diversity).

The topic of indicators on linguistic diversity is also addressed by the Task Group on Measuring the WSIS targets, which is part of the Partnership on Measuring ICT for Development, a multi-stakeholder initiative to improve the availability and quality of ICT data and indicators. This task group will allow feedback in both directions. This task also includes in coordination with W3C/ERCIM, the contribution to various international bodies related to architecture, standardization and normalization on matters concerned with languages, including the working group on multilingualism of the Broadband Commission for Digital Development (http://www.broadbandcommission.org).

Deliverables (brief description) and month of delivery					
Number	Description	WP	Nature	Diss level	Month Due
D4.1	Document of understanding for IP matters	4	R	PU	M6
D4.2	Document of understanding for legal matters	4	R	PU	M6
D4.3	Document of understanding for ethical matters	4	R	PU	M6
D4.4	Linguistic figures version 1 (plus inputs for indicators specs)	4	R	PU	M6
D4.5	Document of recommendations for security matters in WP6	4	R	PU	M12
D4.6	Linguistic figures version 2	4	R	PU	M12
D4.7	Stress test reports of security for WP6	4	R	PU	M24
D4.8	Linguistic figures version 3				M24
D4.9	Strategy document for information society indicators	4	R	PU	M30
D4.10	Strategy document “languages on the Internet” for standard bodies	4	R	PU	M30
D4.11	Linguistic figures final version	4	R	PU	M36

Milestones				
Number	Short description	WP	Month	Means of verification
M4.1	Partner’s meeting for consensus on legal, ethical, privacy and security matters	4	M12	Meeting evaluation
M4.2	Partner’s meeting for finalization on specification for indicators	4	M12	Meeting evaluation
M4.3	Partner’s meeting for strategy decisions for languages matters and information society	4	M36	Meeting evaluation

Work package number	**5**		**Start - End:**		M1 – M36				
Work package title	**SMART SAMPLING OF LARGE-SCALE DATA**									
Activity type	RTD									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	**CNR**	DIALOGIC	CNRS	EXALEAD	UPF	FHG	CWI	FUNREDES
Person-months per participant			50		2	2	20	18	15	
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant	2									

Objectives

The goal of this work package is to give theoretical and practical tools to estimate, as accurately as possible, parameters of the Web ecosystem. Estimations will be done by means of advanced sampling techniques designed specifically for addressing web data and its structure and also used to drive specialized crawling policies.

Tangible outcomes and measures of progress and success

The outcomes of this work package are mainly of two categories. We will design novel sampling techniques for very big and dynamic data. Evaluation of sampling is a well-studied subject. A sample should enable the estimation of indicators of the web page population with low error. That is results should be associated with an error interval and we should not overfit our model on the data we have. Note that an additional assessment of the sampling strategies is performed in Task 12.3. Finally, as another tangible outcome we will develop a set of techniques that will enable the efficient management of the data from which samples and measurements will be drawn.

Description of work

Task 5.1 : Large scale data storage **(M1-M36, 30 PM)**
Task Leader: CNR
Participants: EXALEAD, UPF

This task designs and deploys the large-scale repository for the data-samples collected by the focused crawlers built in WP 5.2 and the indicators specified by WP4.5. Data sources of interest for DILINET are heterogeneous (i.e., they contain different kinds of textual and multimedia data). Moreover, the different modules in the DILINET platform that analyze data, extract and exploit the different Internet indicators designed, must access data locally and very efficiently. Therefore, a scalable and high-performance storage module is strongly needed. The purpose of this task is to design such a large-scale and high-performance infrastructure for storing and accessing the various kinds of DILINET data: raw textual data; audio and video files; indexes built to speed-up data access; knowledge, expressed in the indicators as defined in WP4.5.

Task 5.2 Intelligent sampling strategies **(M1-M33, 37 PM)**

Task Leader: FRAUNHOFER
Participants: CNR, UPF, CNRS, VOCAPIA

This task will develop unbiased and efficient sampling strategies for web pages appropriate for the accurate estimation of statistical indicators of the web. First the target population has to be defined by the partners, with special considerations for cases like dynamic web pages. Existing approaches will be analyzed and algorithms will be developed. The main emphasis will be on unbiasedness and efficiency, e.g. fast coverage of the whole sampling space. To reduce low accuracy for specific indicators a feedback from the analysis work packages may be taken into account which triggers additional sampling, e.g. stratified sampling of specific country code top level domains. Prototypes of the algorithms will be implemented and evaluated on synthetic and real data.

Task 5.3 Intelligent data gathering **(M1-M36, 42 PM)**
Task Leader: CNR
Participants: CWI

Differently from traditional web crawling, in DILINET we do not need to maintain huge document corpora in a repository for future use. On the other hand, DILINET provides to perform large-scale crawling and large-scale data analysis on the data samples retrieved in order to collect important statistics leading to estimating with the highest precision possible the actual values of the indicators devised in in WP4.5. In this task different solutions will be investigated aimed at pushing forward the limits (and backward the costs) of distributed web crawling. Also, volunteer computing will be investigated as a means of granting the huge network and computational bandwidth needed to actually download and analyze large data samples.

Deliverables (brief description) and month of delivery

Number	Description	WP	Nature	Diss level	Month Due
D5.1	Requirements and specifications for the Large scale data storage.	5	R	PU	M6
D5.2	Requirements and specifications for the Intelligent sampling strategies.	5	R	PU	M6
D5.3	Requirements and specifications of the Intelligent data gathering system	5	R	PU	M18
D5.4	Interactions of the Intelligent sampling with the data gathering systems and how data are efficiently stored in the large scale data storage.	5	R	PU	M24
D5.5	Implementation of a prototype of the large scale data storage module.	5	R	PU	M36

Milestones

Number	Short description	WP	Month	Means of verification
M5.1	Specifications of the Large scale data storage architecture	5	M9	Specification document available
M5.2	First implementation of the Large scale data storage including intelligent sampling and data gathering techniques	5	M24	Large scale data storage V1 available
M5.3	Final implementation of the Large scale data storage.	5	M36	Final large scale data storage available

Work package number	**6**		**Start - End:**		M1 – M36				
Work package title	**USER CENTERED MEASUREMENT**									
Activity type	RTD									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	**DIALOGIC**	CNRS	EXALEAD	UPF	FHG	CWI	FUNREDES
Person-months per participant		13	1	29				3	2	1
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant		2	3							

Objectives

The goal of this work package is to measure language behaviour on the internet at the most detailed level: the individual user.

Tangible outcomes and measures of progress and success

For each client, information on which language is being used in which type of application (email, word processing, chatting, social networking) at a particular moment in time (day time, evening, working days, weekends) and a particular frequency

Comparisons across countries, across specific background variables (gender, age, nationality etcetera), and across types of applications (e.g., normal language (word processing) versus chatting language, possibly with email as an intermediate case).

A measure of success is the uptake of the client by users, that is, the growth of the installed base of the client in various countries. The quality of the data that is collected via the client (hence via the nonintrusive automated method) will be validated against data that is collected via an online survey (hence the traditional method).

Description of work

Task 6.1 Input specification: indicators, languages, panel, survey **(M1-M31, 8 PM)**
Task Leader: MAAYA
Participants: NIELSEN, DIALOGIC

This task consists of two parts: developing a protocol and composing a panel. Development of the protocol includes the definition and operationalization of indicators (dependent variables: language use; independent variables: relevant background variables). Two (linked) versions of the protocol will be made: one for the automated data collection and one for the online survey.

The number and type of background variables largely determine the design of the panels. In each country and/or language panel size should be sufficiently big to enable cross-sections that generate statistically significant results. We aim for a minimal panel size of 1,000, in at least ten countries (including at least 3 different continents and 3 emerging economies). This requires an active distribution of the client and a continuous recruiting and selection of new panellists via the international networks of the various partners involved.

Task 6.2 Supported configurations specification **(M1-M3, 5 PM)**
Task Leader: DIALOGIC
Participants: FRAUNHOFER, FUNREDES, NIELSEN

The number and type of devices people use to connect to the Internet is steadily increasing lately. Next to the traditional desk top we now have laptops, netbooks, smartphones, tablets and other types of devices. Most of these devices run on specific versions of operating systems. Although we can re-use some modules in practice this means we will have to develop tailor-made versions of the client for each of the versions that is involved in the project. Hence the scoping of the project is critical to the feasibility of task 6.4. This task is to determine the number of type of operating systems that will eventually be supported. In order to do so, we will draft a cost-benefit analysis in which we will make a sensible trade-off between the benefits of adding another OS and the marginal costs involved in doing so. For instance, Linux is widely used in some emerging economies (China, Brazil, India) and Google's Android is derived from Linux (which probably makes it easier to re-use parts of the source code).
The set of operating systems includes:

- non-mobile a.k.a. desk top and laptop: Windows (7, Vista, XP), Apple OS X (10.7 down to 10.2), Linux (Ubuntu, Fedora, SUSE)
- mobile: iPhone OS (5.*, 4.*), Android (3.*, 2.*), Blackberry (7.*, 6,*, 5.*)

Task 6.3 Panel management **(M13-M31, 14 PM)**
Task Leader: DIALOGIC
Participants: MAAYA, UNESCO

This task covers the initial distribution of the monitoring software (and the subsequent transmission of survey questions, see WP6.5). User panels, of users that accept the installation of the software, will be managed here, via email and an automated multilingual help desk, or bug management system (tickets). A CRM-based system will be used to scrupulously manage the relations with the panellists, which is expected to be highly dynamic (panel churn). The system is needed to target the recruitment of new panellists in WP6.1 (new candidates should match the old ones on background variables), to minimize administrative burden to the panellists (esp. with regard to the online survey), and to manage reminders and messages to dormant users.

Task 6.4 Development of the user-centric measurement client **(M2-M12, 15 PM)**
Task Leader: DIALOGIC
Participants: CNR, FRAUNHOFER, MAAYA

Programming the monitoring client is the heart of this work package. Monitoring user behaviour can be done at various layers of the OSI model (e.g., recording individual key strokes ('keylogging'), in combination with the identification of the current active application, thus recording simultaneous use of different languages in different applications. Identification of the specific user of the device is done automatically (e.g., using Windows API to find the current user name) or by manual self-identification. Identification of the languages used will be based in the technology that is being developed by NUT (in WP8.2). Giving the highly sensitive nature of the data that is collected by the client, maximum care will be given to the privacy of the respondent and the security of the data. All data will be anonymised

locally, at the client side, before it is being sent over an encrypted connection. Only data at a meta-level of the original texts (e.g., type of language, generic characteristics of the text etcetera) will be transferred to and stored at the central server. Consequently, the language recognition module from NUT will run at the client side, not at the server side, and should therefore be extremely light and fast to minimize use of resources from the devices on which the client is being installed.

Task 6.5 Development of the survey module **(M2-M13, 5 PM)**
Task Leader: DIALOGIC
Participants:

The survey module is an integral and important part of the client (WP6.4). Survey questions appear in pop-ups in the client and can be triggered automatically, by specific events, or manually. In the latter case, tailor-made questions are sent to specific subsets of respondents. The identity of the respondent is at no time known of revealed. The ID in our micro data is matched to a specific respondent via hash tables. Thus we can send survey questions to a specific respondent without ever knowing the identity of the respondent. One particular event that always triggers the survey module is the initial installation. After installation of the client some questions will be asked to create a personal profile for each household member (that is, each potential user of the electronic device). Secondly, several questions will be asked to fill the personal profiles. Some of these questions related to the aforementioned background variables (which are usually not known ex ante). Some additional questions will cover the complementary usage of electronic devices (e.g., in most cases the monitoring tool will only be installed on one specific desk top computer at home but many households use multiple computers and devices – also at their work or school – that are not covered by the measurement). To improve the external validity of the measurement results, it is important to know which part of the relevant activities is *not* covered by the measurement.
Use of the survey module is strictly limited to the purposes of this research project. The most important goal of the survey module is to directly collect data from respondents that can be used to validate the data that is automatically being collected about the respondent's behaviour.

Task 6.6 Combined client-survey back end interface **(M4-M18, 7 PM)**
Task Leader: DIALOGIC
Participants: CWI, FRAUNHOFER

Since the online behaviour of hundreds of users is being monitored almost continuously the client generated a large amount of data. For the further development and refinement of the protocol (WP6.1), and the possible re-direction of the research questions during the course of research, it is important to be able to quickly analyse the data. Without a proper user interface it is nearly impossible to analyse the (raw) data in an efficient and timely manner. In this task we will therefore develop a web-based interface that enables a user-friendly access to the data. Furthermore, with the help of the interface the input and output from the client and survey module can be linked on the level of individual (anonymous) respondents or subsets of respondents. The interface also has functionalities to generate tailor-made reports on the fly, and to export those reports to other applications (spreadsheets, statistical packages) for further analysis.

Deliverables (brief description) and month of delivery

Number	Description	WP	Nature	Diss level	Month Due
D6.1	Report on scope of the project in terms of number and types of operating systems supported	6	R	PU	M3
D6.2	First stable version of protocol	6	R	PU	M3

D6.3	Mailing list of panel (initial set of respondents)	6	R	RE	M12
D6.4	Final stable version of protocol	6	R	PU	M22
D6.5	User-centric measurement client, including survey module and back end interface	6	P	RE	M36
D6.6	Final report based on the analysis of the user-centric measurements	6	R	PU	M36

Milestones				
Number	Short description	WP	Month	Means of verification
M6.1	Baseline protocol ready	6.1	M5	Protocol has been thoroughly tested internally and tested successfully on an external panel
M6.2	Baseline panel ready	6.1	M12	A panel has been composed that meets the critical statistical size in each country involved
M6.3	Beta version 1.0 of client ready	6.4	M12	First version of the client available, technically tested (bug free)
M6.4	Beta version 1.1 of client ready	6.5	M13	Second version of the client available, with survey module integrated, technically tested (bug free)
M6.5	Panel active for 9 months (halfway full duration)	6.3	M22	Client installed at devices from 10,000 users in 10 countries, with a consistent 90% of active users
M6.6	Debriefing of panel (end of duration, after 18 months)	6.3	M31	All 10,000 users informed about the completion of the measurement.

Work package number	**7**		**Start - End:**		M1 – M36				
Work package title	**LANGUAGE INDICATORS IN AUDIO**								
Activity type	RTD								
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	DIALOGIC	**CNRS**	EXALEAD	UPF	FHG	CWI	FUNREDES
Person-months per participant		2			39			2	2	
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant	50		8							

Objectives

Much of the information on the web is not in a textual format, and therefore will escape detection, classification and categorization via text-based methods. This work package will be concerned with sampling a wide range of websites hosting audio and video documents, and developing methods to identify the languages spoken in them. A second activity will be to explore techniques to identify the content categories and other information in the audio channel.

Tangible outcomes and measures of progress and success

Prototype language identification systems will be developed and made accessible to the partners via a Web-based service. Performance will be assessed on representative test data selected by the partners. Content categories will be annotated in the automatic transcripts for languages which STT systems are available (task 7.4) . The automatic transcripts provided via the transcription service will also be used as text input to allow the developed opinion and sentiment analysis methods of WPX to be applied to audio and audiovisual documents.

We will also measure progress and success by taking part in appropriate international campaigns organized outside of the project such as the Language Recognition benchmarks organized by the NIST (the US National Institute of Standards and Technology, www.itl.nist.gov/iad/mig/tests/lre).

Description of work

Task 7.1 Specification of the representative corpus obtained via large scale sampling **(M1-M6, 13 PM)**
Task Leader: VOCAPIA
Participants: CNRS, NIELSEN, CWI, FRAUNHOFER

In collaboration with the work in work package 5, a large sample of heterogeneous audio and audiovisual corpora will be identified.

Task 7.2 Annotation of the representative corpus with language tags **(M1-M15, 18 PM)**

Task Leader: CNRS
Participants: VOCAPIA, NIELSEN

This task is concerned with the annotation of a corpus annotated with language tags. This corpus is required for both model training and for evaluation purposes. Innovative methods will be explored to obtain the labelled data by incorporating speech technologies in the annotation process.
For example, since a single audio document may contain segments in different languages, the document will first be automatically partitioned into clusters segments for each language, and individual clusters will be presented to humans for annotation.

Task 7.3: Language identification from audio **(M1-M36, 36 PM)**
Task Leader: VOCAPIA
Participants: CNRS

This task is concerned with developing models for language identification, that is the task of automatically determining the language of a given speech segment. The general task of language recognition can be divided into several sub-tasks including language identification and language detection, and can be applied to a fixed or open set of languages.

The performance of the widely used phonotactic approaches is highly dependent on the quality of the underlying phone decoders. Therefore one direction will be the construction of accurate language-specific (L) and language-independent (LI) phone recognizers in order to generate more consistent phone n-gram statistics. Several techniques that have been widely adopted in speech recognition have not yet been investigated for language identification. These include discriminant training for the component acoustic phone models; methods to improve the quality of target language phonotactic models such as optimization of decoding parameters and intelligent selection of phone contexts; and improved decoding making use of multiple hypotheses and automatic learning techniques.

Today's state-of-the-art language identification (LID) systems typically assume that an audio document is only in a single language. This assumption is not always valid, in particular when there is simultaneous translation of audio segments containing speech in a language other than the main language of the document. To address this LID will be compared on fixed sized on X second chunks and on clustered segments found by an automatic partitioner. Accent/dialect identification modules will be developed for highly represented variants.

Task 7.4: Model Development for Automatic Content Category Identification **(M1-M36, 34 PM)**
Task Leader: CNRS
Participants: VOCAPIA

This task is concerned with developing models for automatic content (topic) identification in audio data based on automatic speech-to-text transcription for covered languages. Unsupervised methods will be employed to build baseline speech-to-text transcription systems for most common N languages not covered in order extend the range of languages to which content analysis can be applied. The set of topics will be defined in collaboration with the other partners and in close coordination with WP8.3 concerned with content analysis of text data.

Deliverables (brief description) and month of delivery

Number	Description	WP	Nature	Diss level	Month Due
D7.1	Corpus Specification	7	R	RE	M3

D7.2	Report on Annotated Corpus	7	R	RE	M12
D7.3	Report on LID systems v1	7	R	RE	M6
D7.4	Report on Topic Detection in audio v1	7	R	RE	M14
D7.5	Report on LID systems v2	7	R	PU	M36
D7.6	Report on Topic Detection in audio v2	7	R	PU	M36

Milestones				
Number	Short description	WP	Month	Means of verification
M7.1	Baseline web-based LID	7	M6	service available
M7.2	Baseline topic detection web-based	7	M12	service available
M7.3	Extended web-based LID	7	M18	service available
M7.4	Final web-based LID	7	M34	service available
M7.5	Final topic detection web-based	7	M34	service available

Work package number	**8**		**Start - End:**		M1 – M36				
Work package title	**VALIDITY & ANALYSIS**									
Activity type	RTD									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	DIALOGIC	CNRS	**EXALEAD**	UPF	FHG	CWI	FUNREDES
Person-months per participant		1	10		12	12	6	36		
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant	8		10							

Objectives

This work package develops the tools for identifying valid web pages, cleaning these web pages, and providing a first round of annotation of the web pages.

This Work Package interacts with WP5 which develops the sampling techniques for choosing representative web pages, and WP7 which provides automatic transcripts of speech content. In this WP8, techniques are developed to recognize and eliminate spam, spam farms, duplicate and near duplicate pages. Techniques will also be developed for extracting from the remaining, validated pages, the substantive content of the pages (that is, removing advertisements, header, and footer, and menus). From this clean content, other research and development activities in this work package will perform language recognition, gross category recognition (for example, medicine, economics, sports), opinion detection with respect to document content or single concepts (e.g. negative, neutral, positive), and extract information about outgoing links that will be stored and analyzed to provide inter-language linkage indicators (such as what percentage of Spanish pages link to Catalan).

The results of this work package will be exploited by the deeper content analysis task of WP9: “DATA & KNOWLEDGE REPRESENTATION” that will produce the linguistic indicators and language models from this “cleaned” data.

Tangible outcomes and measures of progress and success

Performance of category models and opinion mining algorithms will be evaluated on representative test corpora selected by the partners. If possible human gold standard annotations will be used for testing; otherwise automatically translated text with annotations will be employed to assess performance. We will also measure performance and applicability by taking part in appropriate international comparisons organized by initiatives such as CLEF (Cross-Language Evaluation Forum). Tangible outcomes will be web page cleaning programs, modules for topic/language/sentiment identification, and community identification software.

Description of work

Task 8.1 Content cleaning: spam, duplication **(M3-M24, 6 PM)**
Task Leader: EXALEAD
Participants:

This task consists of identifying web pages that are most likely spam. Exalead will develop new language models developed for likely spam pages (pornography, gambling), and package these models into a module that will take a URL in input and decide whether it should be retained in the sample. EXALEAD will also develop techniques for removing duplicate pages or almost duplicate pages from a collection of URLS. In this task, tools will be packaged for extracting the content of a textual content of a page, filtering out structural menus, advertisements and other content-like but non-substantive text.

Task 8.2 Content analysis: language identification **(M3-M6, 11 PM)**
Task Leader: CNR
Participants: MAAYA

This task has to do with developing a tool (implemented as a self-contained Java jar library) that, starting from a clean web page (i.e., a web page that has underwent a cleaning pass by the tool developed in Task 8.1), identifies the language it is written into. The approach that ISTI-CNR will follow in order to develop this tool will be based on single-label classification via supervised learning.
In order to guarantee the computational efficiency of the resulting tool, hierarchical classification technology will be adopted, which will allow exponential savings at both training and classification time with respect to standard, flat approaches. The classification hierarchy will be generated so that languages that are morphologically “close” (e.g., Italian, Spanish, Portuguese) end up in the same sub-hierarchy. In order to bring about robustness to ill-formed language, feature sets based on character n-grams will be used. We will also apply information-theoretic feature selection so as to retain, at each internal node of the hierarchy, only the most discriminative n-grams, thereby ensuring accuracy and efficiency at the same time.

In order to support integration of the library into the overall system, Javadoc documentation and well-documented code samples of its usage will also be provided.

Task 8.3 Content analysis: category recognition and opinion mining **(M1-M36, 66 PM)**
Task Leader: FRAUNHOFER
Participants: CNRS, NIELSEN, VOCAPIA

This task will first associate content categories to a cleaned web page. The partners participating in this task will select a hierarchy of categories (such as FINANCE, SPORTS, MEDICINE) possibly using open multilingual resources like Wikipedia. Then Fraunhofer will develop a category identification model in this space which will assign one or more categories to each web page in all languages treated in the DILINET project. The sense mapping of words between different languages will be supported by unsupervised models, e.g. multilingual topic models.

A second target of this task is the multilingual detection of subjective opinions expressed in web pages. Using resources like dictionaries, automatic translation and linked open data two types of opinion mining techniques will be developed by Fraunhofer. The first technique will target text sections like paragraphs or whole documents which have been assigned to content categories and estimate their overall subjective orientation (e.g. negative, neutral, positive). The second technique will train models to detect opinions with respect to a selected set of concepts represented by phrases (e.g. "Euro",

"Climate Change"). The set of these concepts will be compiled by the partners.

Vocapia and CNRS will develop modules for automatic opinion/sentiment analysis in audio/audiovisual data. In addition to applying text based indicators to automatic transcripts when available, prosodic related features will be extracted from the audio and fused with the text based features (whenever available).

Task 8.4 Network-links analysis **(M12-M24, 12 PM)**
Task Leader: EXALEAD
Participants: UPF

Exalead will develop a module that performs an analysis of a large graph of web sites (the collection sampled) and analyze the graph structure according to a set of attributes developed in WP3 such as country origin, language, content category, and produce an analysis of what is connected to what according to these attributes. UPF will develop tools to detect communities (coherent subgraphs) in the web graph.

Deliverables (brief description) and month of delivery

Number	Description	WP	Nature	Diss level	Month Due
D8.1	First requirements for category analysis	8	R	PU	M6
D8.2	First prototype for category analysis	8	P	RE	M12
D8.3	Updated requirements for category and opinion analysis ...	8	R	PU	M18
D8.4	Second prototype for content and opinion analysis	8	P	RE	M24
D8.5	Updated requirements for content and opinion analysis	8	R	PU	M30
D8.6	Third prototype for content and opinion analysis	8	P	RE	M36

Milestones

Number	Short description	WP	Month	Means of verification
M8.1	First prototype of category analysis	8	M12	software evaluation and test data
M8.2	Second prototype of category and opinion analysis	8	M24	software evaluation and test data
M8.3	Second prototype of category and opinion analysis	8	M36	software evaluation and test data

Work package number	**9**		**Start - End:**		M1 – M36				
Work package title	**DATA & KNOWLEDGE REPRESENTATION**									
Activity type	RTD									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	DIALOGIC	CNRS	EXALEAD	UPF	FHG	**CWI**	FUNREDES
Person-months per participant						4			45	3
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant										

Objectives

WP9 develops the tools to transform the cleaned up and annotated web pages resulting from WP8 tools into representations that provide insight and can be shared with the public. Hereto, we develop a web analytics tool-suite, ranging from low level language modelling to the creation of high level summaries oriented towards the analysts' usage. Exploratory user studies help to iteratively refine the indicators of WP3 and inform (and where needed redirect) the research in WP9.

Description of work

Task 9.1 Conceptual maps **(M6-M30, 12 PM)**
Task Leader: CWI
Participants:

Develop the tools to summarize the structure of crawled data at a more aggregate level to extract more meaning more easily. Apply text mining methods to create (domain-specific) thesauri or taxonomies, per language or per topic (grouped by language and/or region). Concrete outcomes of this task enable users to analyse the perception of events and facts per region and/or language.

Task 9.2 Query language and support over language models **(M1-M36, 18 PM)**
Task Leader: CWI
Participants:

Develop an easy-to-use interactive tool that will allow the user to specify their needs and transform the crawled data in context of their task, through an underlying 'query language' that expresses operations over the types of (semi-structured, aspect, or context-aware) statistical language modelling techniques studied in task 9.3.

Task 9.3 Language models **(M12-M36, 10 PM)**
Task Leader: EXALEAD
Participants: CWI

Define new statistical models of languages and sublanguages and enhance statistical language modelling techniques, specifically in creating baseline language models for the general languages treated in DILINET, as well as for topic specific subsets of language. Explore baseline models for trend analysis and novelty measurement.

Task 9.4 Correlation relevance and derivation of new indicators **(M12-M36, 12 PM)**
Task Leader: FUNREDES
Participants: CWI

Set up exploratory user studies to both measure the impact of the output of tasks 9.1 – 9.3 from an end-user perspective and provide evaluation results as the basis of the next iteration of tool development. Outcomes of this task allow assessing the correlation between data and meaning representations and language models of the web content on the one side and given indicators on the other side, instrumental for the derivation of new indicators.

Deliverables (brief description) and month of delivery

Number	Description	WP	Nature	Diss level	Month Due
D9.1	Query Language Specification (T9.2)	9	R	PU	M9
D9.2	Tool for Conceptual Maps V1 (T9.1)	9	P	RE	M12
D9.3	User Study – Conceptual Maps (T9.4)	9	R	PU	M18
D9.4	Language Models Publicly Available (T9.3)	9	R	PU	M30
D9.5	Interactive Tool (T9.2)	9	P	RE	M24
D9.6	User Study – Conceptual Maps with topics/opinions (T9.2 and T9.4)	9	R	PU	M30
D9.7	Tool for Conceptual Maps V2 (T9.1, using T9.2)	9	P	RE	M36

Milestones

Number	Short description	WP	Month	Means of verification
M9.1	Pilot of first user study completed	9	M15	results of pilot study available
M9.2	Language Models Prepared for Internal Use	9	M24	A demo application based on statistical language modelling, e.g. word breaking, uses the DILINET models
M9.3	Pilot of second user study completed	9	M27	results of pilot study available

Work package number	**10**		**Start - End:**		M1 – M36				
Work package title	**SYSTEM DEVELOPMENT & INTEGRATION**									
Activity type	RTD									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	DIALOGIC	CNRS	EXALEAD	UPF	**FHG**	CWI	FUNREDES
Person-months per participant		5	22.8	7.8	5.8	2	3.8	45.3	3.8	
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant	3.8		3.3							

Objectives

This work package provides a framework for easy integration of the results of the other work packages in runnable prototypes.
The completion of the different project prototypes will be coordinated and supported by providing the infrastructure and common tools.

Tangible outcomes and measures of progress and success
The main outcomes of this WP are the three DILINET prototypes. The requirements of these prototypes are specified in the System Requirements and Architecture Reports available six months before the prototype releases. Manual inspection of software modules as well automatic software tests and metrics will be employed to generate an up-to-date assessment of functional and non-functional adequacy as well as the software quality of the different software modules provided by the different work packages.

Description of work

Task 10.1 System requirements **(M1-M30, 19 PM)**
Task Leader: FRAUNHOFER
Participants: CNR, DIALOGIC, CNRS, EXALEAD, UPF, CWI, VOCAPIA, NIELSEN

The large-scale DILINET modules motivate a number of requirements for the development of DILINET tools, services and overall system. The aim of this task is to analyze and specify the functional and non-functional systems requirements arising from work packages 5 to 9 and WP11. The functional requirements describe the DILINET capabilities provided and required by the work packages for solving a functional, application-specific problem. The non-functional requirements refer to system requirements which are not of a functional nature, but contribute decisively to the applicability of the system. They define, e.g. quality requirements, safety and security requirements or performance requirements.
All partners specify the requirements for their software modules. Based on these requirements the requirements for the tools, services and overall system are derived.

Task 10.2 Terminology definition **(M1-M6, 5.4 PM)**
Task Leader: MAAYA
Participants: CNR, DIALOGIC, CNRS, UPF, FRAUNHOFER, CWI, VOCAPIA, NIELSEN

The participants of the DILINET project have a very diverse professional background. To allow an unambiguous communication this task will establish an agreed set of common concepts which will be defined in a glossary. This common terminology has to cover all technical and non-technical aspects of the DILINET project and its targeted applications. The glossary will be produced in English as well as a limited set of languages which will be defined during the first stage of the project in coordination with involved partners.

Task 10.3 System architecture **(M1-M30, 24 PM)**
Task Leader: FRAUNHOFER
Participants: CNR, DIALOGIC, CNRS, UPF, CWI, VOCAPIA, NIELSEN

The main objective of this task is to define a detailed architecture design of the DILINET system. The design will identify the components and services that will be implemented by the project. The system architecture design will also specify the main selected technologies, protocols, and middleware that will be used in the project.

Task 10.4 Project infrastructure, process and tools **(M1-M36, 30 PM)**
Task Leader: FRAUNHOFER
Participants: CNR

This task will implement all infrastructure elements of DILINET, including interfaces, components and services allowing plugging and interoperation of external (legacy systems) and internal systems, tools and services. The actual DILINET components developed in work packages 5-9 will be responsible for the development of all services, interfaces, and processes defined in this work package.

Task 10.5 Prototype development cycle **(M1-M36, 25 PM)**
Task Leader: FRAUNHOFER
Participants: MAAYA, CNR, DIALOGIC, CNRS, EXALEAD, UPF, CWI, VOCAPIA, NIELSEN

For the different prototypes provided by DILINET which are increasingly comprehensive, this task will monitor the compliance of modules to the standards defined in Tasks 10.1-10.4. The task will also check the adequacy of software tests and optimizations. Finally this task will update requirements arising during the course of the project in a concerted way and adapt the project terminology, infrastructure, processes, and tools accordingly.

Deliverables (brief description) and month of delivery					
Number	Description	WP	Nature	Diss level	Month Due
D10.1	Terminology Definition Report	10	R	PU	M02
D10.2	System Requirements Report	10	R	PU	M04
D10.3	First System Architecture	10	R	PU	M06
D10.4	First Prototype	10	P	RE	M12
D10.5	Updated System Requirements and Architecture Report	10	R	PU	M18
D10.6	Second Prototype	10	P	RE	M24

D10.7	Final System Requirements and Architecture Report	10	R	PU	M30
D10.8	Third Prototype	10	P	RE	M36

Milestones				
Number	Short description	WP	Month	Means of verification
M10.1	First Prototype	10	M12	Automatic software tests and metrics as well as manual inspection according to requirements.
M10.2	Second Prototype	10	M24	Automatic software tests and metrics as well as manual inspection according to requirements.
M10.3	Third prototype	10	M36	Automatic software tests and metrics as well as manual inspection according to requirements.

Work package number	**11**		**Start - End:**		M1 – M36				
Work package title	**APPLICATIONS**									
Activity type	RTD									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	DIALOGIC	CNRS	EXALEAD	UPF	FHG	CWI	FUNREDES
Person-months per participant		20		1	4	12		6		
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	**NIELSEN**							
Person-months per participant	6	2	22							

Objectives

Various applications are implemented, under a common methodological framework, in order to demonstrate, in concrete and live environments, the usefulness of the developed research and obtained products. The applications focus three main lines of activity in order to cover the scope of the project research and results: intelligent data management as related to search engines index creation, language technologies in the digital world and linguistic diversity indicators as tool for business in the digital economy and as input for public languages and information society policies. From the applications evaluation future impacts are forecasted and documented,

Tangible outcomes and measures of progress and success
The main output of this WP is key performance indicators as volume metrics, opinion driving indicators, opinion/sentiment and context information. These indicators will provide to public policy and enterprises quantitative and qualitative information of cultural attitudes of the society which enable public policy and enterprises to communicate and develop services in a better way. The effectiveness and efficiency of this WP will be measured based on tracking the access to services by public policy and enterprises as well as measuring the success of using the information out of this WP for decision making process within public policy and enterprises. This will be supported by case studies.

Description of work

Task 11.1 Applications evaluation and future impacts forecast **(M24-M36, 8 PM)**
Task Leader: MAAYA
Participants: DIALOGIC, VOCAPIA

The activities consist in the creation of a methodological framework for the applications, to both, evaluate the impacts obtained by the research which have been conducted as well as the results which have been obtained and forecast future impacts of the project. In that sense, T11.1 will act both as front-end and back-end to the following tasks of WP11 and play a transversal role to help emerge coherence and synergy from the different parts of the projects viewed from the users and applications point of view. Research impacts are foreseen in various directions, from the creation of new methods

for crawling to progress in language technologies and including the generalization of user centered approaches for information society indicators and the first explorations of new domains such as the content characterization or the social networks dynamics as well as the use of conceptual maps or visual analytics for fields which are different but can be derived from linguistic diversity. The reports produced by T11.1, together with those of T3.4 and T3.5, will provide valuable insights for future research and applications driven from the DILINET project as well as present a comprehensive and coherent balance of the outcomes of the project as they impact the various fields involved (intelligent data management, voice technologies, linguistic diversity, information society indicators, content industry, public policies for digital divide…).

Task 11.2 Pilot for intelligent information management **(M1-M36, 16 PM)**
Task Leader: EXALEAD
Participants: CNRS, VOCAPIA

DILINET will provide a means of determining which parts of the Web are currently under indexed in the major search engines. This application will create a local search engine for two EU countries that have been judged to be underrepresented. The application will crawl those countries, particularly the sites that have been identified as non-spam but not present in the major Web search engines. For each of the crawled countries, a demonstrator will be built that shows what the index of that country can be if adequately indexed.

Task 11.3 Pilot for linguistic diversity public policy **(M1-M36, 12 PM)**
Task Leader: MAAYA
Participants: UNESCO

This application has been defined in two complementary pieces which will apply the DILINET research results into concrete case study related to digital libraries and huge international websites. The first part is applied to a selection of he most important digital libraries with non national only coverage (Google Book, Europeana, ScholarVox, World Digital Library). It consists in both the search of duplicated records and the identification of contents by language maintaining a permanent watch in order to capture the trends. This will offer a tool for management of libraries.
The second part is applied into a selection of websites of important international organisations, primarily to the EC sites, in order to evaluate the language repartition of their content and monitor the evolution. The selection is composed first by the sites of European Union in order to have a feed-back of how the member states languages are represented and is completed by a selection of United Nation's agencies made upon a multilingual criteria for the use of official languages (such as UNICEF or WHO).

Task 11.4 Pilot for linguistic diversity business **(M1-M36, 37 PM)**
Task Leader: NIELSEN
Participants: FRAUNHOFER, MAAYA, CNRS, VOCAPIA

This application will use the DILINET s data and evaluation methods for the identification of actual trends and change of language for industry. Within this application the research data of DILINET will be translated into relevant key performance indicators including opinion/sentiment that enterprises can use for developing, promoting and protecting their content, products, services etc; this application will also support SMEs (small and medium enterprises) to identify niches and new trends for new innovation. This application will provide volume metrics, opinion driving indicators, opinion/sentiment (good, bad, neutral) on specific concepts and context information (interrelation between language and/or opinion/sentiment information). One further objective of this application is to transfer information into graphs and diagrams.
Nielsen together with the partners will specify an appropriate task. Then the partners will collect or

provide the required data and Fraunhofer and CNRS will adapt their analysis procedures and apply them for content analysis and opinion mining on text (and audio?). Nielsen will supervise the work, evaluate the results and provide the results to pilot customers.
Furthermore the success of the analysis will be measured by effectiveness and efficiency studies to get insights in the value of DILINET data and results within the innovation and decision making process of enterprises as well as the economic success of content, products, services etc. created and developed with DILINET data.

Deliverables (brief description) and month of delivery					
Number	Description	WP	Nature	Diss level	Month Due
D11.1	Applications evaluation framework	11	R	PU	M24
D11.2	Intermediary Report from intelligent information management pilot	11	R	PU	M30
D11.3	Intermediary Report from linguistic diversity public policy pilot	11	R	PU	M30
D11.4	Intermediary Report from for linguistic diversity business pilot	11	R	PU	M30
D11.5	Final Report from intelligent information management pilot	11	R	PU	M36
D11.6	Final Report from linguistic diversity public policy pilot	11	R	PU	M36
D11.7	Final Report from for linguistic diversity business pilot	11	R	PU	M36
D11.8	Impacts forecast Report	11	R	PU	M36

Milestones				
Number	Short description	WP	Month	Means of verification
M11.1	Consensuated Evaluation Framework and launch of Applications	11	24	Partner meeting
M11.2	Intermediary evaluation and reporting	11	30	Advisory Boards evaluation of reports
M11.3	Final reporting	11	36	Final meeting with presence of Advisory Boards

Work package number	**12**		**Start - End:**		M1 – M36				
Work package title	**RESULTS ASSESSMENT/EVALUATION**									
Activity type	RTD									
Participant number	1	2	3	4	5	6	7	8	9	10
Participant short name	ERCIM	MAAYA	CNR	DIALOGIC	CNRS	EXALEAD	UPF	FHG	CWI	**FUNREDES**
Person-months per participant		14		3				3		18
Participant number	11	12	13	14	15	16	17	18	19	20
Participant short name	VOCAPIA	UNESCO	NIELSEN							
Person-months per participant										

Objectives

This work-package have the transversal function to maintain a constant focus on the desired products and offer the tools to assess both the usefulness and impact potential of the new methods as well as the obtained products and provide fast feedback on the research part in order to provide efficient guidance. It will offer to project management an informed decision framework from the point of view of the very objective of the project and will serve as a central service for the evaluation and validation of the indicators data produced by various line of research watching carefully for divergence in the production. For that reason it will also play a special role in reducing the natural trend of different line of research to work stand-alone and be a source of cross-discipline synergy seeking. The responsibility of this work-package, in constant negotiation with project management, is to be the guarantee that the investment leads to tangible and valid results and to the best cross-discipline standards.

Tangible outcomes and measures of progress and success

The tangible outcomes include:

- an adaptation of the “words sampling” method to work with crawling instead of search engines and which will be verified in a simple and well defined set of web pages;
- the production of results from LOP and “words sampling” methods to be used as reference data to contrast new methods and results;
- a thorough process of assessment of the products of the project in terms of usefulness and potential impact, from the users point of view;
- a thorough process of testing and internal validation of the project as a system;
- a thorough process of testing and external validation of the project from the methodology standpoint.

All those processes will be consolidated into comprehensive reports which will be evaluated by the Advisory Boards.

Description of work

Task 12.1 Internal validity: results crosschecking and comparisons **(M1-M36, 10 PM)**
Task Leader: FUNREDES
Participants: MAAYA

This is a central activity to assess the results of the indicators production and crosscheck the results for validity. Both context driven and statistical monitoring methods will be developed together for result's bias evaluation and control. This task also stands on the data produced by the proven methodologies used in task 12.2 and 12.3.

Task 12.2 Recognition algorithm method **(M1-M36, 7 PM)**
Task Leader: MAAYA
Participants:

The activity consists in adapting and applying language recognition algorithms and the existing LOP project (managed by NUT) with the new elaborated strategies for crawling as produced by WP5. The product will serve as input to 12.1 to assess the new methods produced by WP5.

Task 12.3 Words sampling method **(M1-M36, 7 PM)**
Task Leader: FUNREDES
Participants: MAAYA

The activity consists in adapting and applying the words sampling method of FUNREDES/UNION LATINE with the new elaborated strategies for crawling as produced by WP5. The product will serve as input to 12.1 to assess new methods produced by WP5.

Task 12.4 Utility and Impact assessment **(M1-M36, 8 PM)**
Task Leader: MAAYA
Participants: FUNREDES

The activity consists of the fast assessment of the intermediary methods and results from the research components of the project from the point of view of usefulness and potential for impact. The usability of the project results will be periodically checked against a set of indicators that will take into account both the technological and societal aspects of the DILINET work.

Task 12.5 System testing and validation **(M7-M36, 3 PM)**
Task Leader: FRAUNHOFER
Participants:

In this task the overall DILINET system will be tested and evaluated to provide stable, fully integrated, optimized and tested versions of DILINET. While task 10.5 concentrates on individual software components this task considers the overall system, which will be assessed with respect to the global adequacy of the developed software

Task 12.6 External validity – methodological soundness **(M1-M36, 3 PM)**
Task Leader: DIALOGIC
Participants:

The activity consists on a global audit of the external validity of the research results. The base hypothesis is that these results can widely be made general. However several threats to external validity might occur when independent variables depend on factors that are not explicitly included in

the research design. Such factors include the research objects (sampling of websites and panellists), the researchers themselves, the location (country, urban/rural split within a country), and the setting (home/office, devices that are also used by the panellist but that are not being monitored). We will scrutinize the possible occurrence of all factors and also keep an eye on possible trade-off with internal validity ('ecological validity').

Deliverables (brief description) and month of delivery					
Number	Description	WP	Nature	Diss level	Month Due
D12.1	Software development to adapt FUNREDES/UL method to crawling	12	P	RE	M12
D12.2	Reference measurement made with FUNREDES/UL method	12	R	PU	M18
D12.3	Reference measurement made with LOP method	12	R	PU	M18
D12.4	Result analysis and assessment first report	12	R	PU	M18
D12.5	System validation first report	12	R	PU	M18
D12.6	Methodology assessment first report	12	R	PU	M18
D12.7	Result analysis and assessment second report	12	R	PU	M24
D12.8	System validation second report	12	R	PU	M24
D12.9	Methodology assessment second report	12	R	PU	M24

Milestones				
Number	Short description	WP	Month	Means of verification
M12.1	FUNREDES/UL method adaptation	12	M12	Software test in sample data.
M12.2	Reference measurements obtained	12	M18	Crosschecking between T12.2 and T12.3
M12.3	Assessment first reports	12	M18	Partner meeting
M12.4	Assessment second reports	12	M24	Partner meeting

Summary of effort

The following table provides a summary of the planned effort, for each work package by each participant in person-months. The work-package leader for each WP is identified by showing the relevant person-month figure **in black**.

Partic. no.	Partic. short name	WP 1	WP 2	WP 3	WP 4	WP 5	WP 6	WP 7	WP 8	WP 9	WP 10	WP 11	WP12	Total PM per Partner
1	ERCIM	**22**		5										27.0
2	MAAYA	1		**15**	**25**		13		1		5	20	14	94.0
3	CNR			1		**50**	1		10		23			84.8
4	DIALOGIC			2	1		**29**				8	1	3	43.8
5	CNRS			4		2		**39**	12		6	4		66.8
6	EXALEAD		**13**	2		2			**12**	4	2	12		47.0
7	UPF		4	4		20			6		4			37.8
8	FRAUNHOFER			1		18	3	2	36		**45**	6	3	114.3
9	CWI			1		15	2	2		**45**	4			68.8
10	FUNREDES	3		8	12		1			3			**18**	45.0
11	VOCAPIA			3		2		50	8		4	6		72.8
12	UNESCO			15	6		2					2		25.0
13	NIELSEN			2			3	8	10		3	**22**		48.3
Total PM per WP		26.0	17.0	63.0	44.0	109.0	54.0	101.0	95.0	52.0	103.4	73.0	38.0	775.4

Figure 12 - Summary of effort at WP level

The following table gives a more detailed overview of the planned effort, with each WP broken down in tasks and giving the contribution by each participant in person-months.

			1	2	3	4	5	6	7	8	9	10	11	12	13		
			ERCIM	MAAYA	CNR	DIALOGIC	CNRS	EXALEAD	UPF	FRAUNHOFER	CWI	FUNREDES	VOCAPIA	UNESCO	NIELSEN		
	Note: PMs/WP for WP leaders indicated in bold/red	Leader	-	-	-	-	-	-	-	-	-	-	-	-	-		
WP 1	**Project Management**	**ERCIM**	**22.0**	**1.0**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**3.0**	**-**	**-**	**-**	**26.0**	**3.4%**
T1.1	Co-ordination of activities	ERCIM	8.0													8.0	1.0%
T1.2	Resource management	ERCIM	3.0													3.0	0.4%
T1.3	Project reporting	ERCIM	4.0													4.0	0.5%
T1.4	Deliverable quality assurance	MAAYA	2.0	1.0												3.0	0.4%
T1.5	Liaison with the EC	ERCIM	3.0													3.0	0.4%
T1.6	Collaborative tools	ERCIM	2.0									3.0				5.0	0.6%
WP 2	**Scientific Coordination**	**EXALEAD**	**-**	**-**	**-**	**-**	**-**	**13.0**	**4.0**	**-**	**-**	**-**	**-**	**-**	**-**	**17.0**	**2.2%**
T2.1	Scientific and Technical Coordination	EXALEAD						8.0	1.0							9.0	1.2%
T2.2	Project Technical Assessment	EXALEAD						3.0	1.0							4.0	0.5%
T2.3	Update of the description of tasks	EXALEAD						1.0	1.0							2.0	0.3%
T2.4	Board Chairing	EXALEAD						1.0	1.0							2.0	0.3%
																-	0.0%
																-	0.0%
WP 3	**Dissemination & Exploitation**	**MAAYA**	**5.0**	**15.0**	**1.0**	**2.0**	**4.0**	**2.0**	**4.0**	**1.0**	**1.0**	**8.0**	**3.0**	**15.0**	**2.0**	**63.0**	**8.1%**
T3.1	Research communication & dissemination	UPF	3.0		1.0	1.0	4.0		2.0	1.0	0.5		1.0			13.5	1.7%
T3.2	Business communication & dissemination	EXALEAD	2.0			1.0		2.0			0.5		2.0		2.0	9.5	1.2%
T3.3	Indicators communication & dissemination / linguistic policies	MAAYA		9.0										9.0		18.0	2.3%
T3.4	Indicator production sustainability	MAAYA		5.0								5.0		6.0		16.0	2.1%
T3.5	System/technologies maturity roadmap	UPF		1.0					2.0			3.0				6.0	0.8%
																-	0.0%
WP 4	**Societal Issues**	**MAAYA**	**-**	**25.0**	**-**	**1.0**	**-**	**-**	**-**	**-**	**-**	**12.0**	**-**	**6.0**	**-**	**44.0**	**5.7%**
T4.1	Intellectual property matters	MAAYA		3.0												3.0	0.4%
T4.2	Legal, ethical and regulatory matters	MAAYA		3.0										3.0		6.0	0.8%
T4.3	Privacy protection and security technologies	MAAYA		5.0		1.0										6.0	0.8%
T4.4	Linguistic diversity figures inventory, watch and stocktaking	MAAYA		13.0								6.0				19.0	2.5%
T4.5	Information Society indicators matters	FUNREDES		1.0								6.0		3.0		10.0	1.3%
WP 5	**Smart Sampling of Large-Scale Data**	**CNR**	**-**	**-**	**50.0**	**-**	**2.0**	**2.0**	**20.0**	**18.0**	**15.0**	**-**	**2.0**	**-**	**-**	**109.0**	**14.1%**
T5.1	Large scale data storage	CNR			20.0			2.0	8.0							30.0	3.9%
T5.2	Intelligent sampling strategies	FRAUNHOFER			3.0		2.0		12.0	18.0			2.0			37.0	4.8%
T5.3	Intelligent data gathering	CNR			27.0						15.0					42.0	5.4%
																-	0.0%
																-	0.0%
																-	0.0%
WP 6	**User Centered Measurement**	**DIALOGIC**	**-**	**13.0**	**1.0**	**29.0**	**-**	**-**	**-**	**3.0**	**2.0**	**1.0**	**-**	**2.0**	**3.0**	**54.0**	**7.0%**
T6.1	Input specification: indicators, languages. panel, survey	MAAYA		6.0		1.0									1.0	8.0	1.0%
T6.2	Supported configurations specification	DIALOGIC				1.0				1.0		1.0			2.0	5.0	0.6%
T6.3	Panel management	DIALOGIC		6.0		6.0								2.0		14.0	1.8%
T6.4	Development of the user-centric measurement client	DIALOGIC		1.0	1.0	12.0				1.0						15.0	1.9%
T6.5	Development of the survey module	DIALOGIC				5.0										5.0	0.6%
T6.6	Combined client-survey back end interface	DIALOGIC				4.0				1.0	2.0					7.0	0.9%
WP 7	**Language Indicators in Non-Text Media**	**CNRS**	**-**	**-**	**-**	**-**	**39.0**	**-**	**-**	**2.0**	**2.0**	**-**	**50.0**	**-**	**8.0**	**101.0**	**13.0%**
T7.1	Specification of the representative corpus obtained via large scale samp	VOCAPIA					2.0			2.0	2.0		4.0		3.0	13.0	1.7%
T7.2	Annotation of the representative corpus with language tags	CNRS					9.0						4.0		5.0	18.0	2.3%
T7.3	Language identification from audio	VOCAPIA					12.0						24.0			36.0	4.6%
T7.4	Model Development for Automatic Content Category Identification	CNRS					16.0						18.0			34.0	4.4%
																-	0.0%
																-	0.0%
WP 8	**Validity & Analysis**	**EXALEAD**	**-**	**1.0**	**10.0**	**-**	**12.0**	**12.0**	**6.0**	**36.0**	**-**	**-**	**8.0**	**-**	**10.0**	**95.0**	**12.3%**
T8.1	Content cleaning: spam, duplication	EXALEAD						6.0								6.0	0.8%
T8.2	Content analysis: language recognition	CNR		1.0	10.0											11.0	1.4%
T8.3	Content analysis: topic recognition & opinion mining	FRAUNHOFER					12.0			36.0			8.0		10.0	66.0	8.5%
T8.4	Network-links analysis	EXALEAD						6.0	6.0							12.0	1.5%
																-	0.0%
																-	0.0%
WP 9	**Data & Knowledge Representation**	**CWI**	**-**	**-**	**-**	**-**	**-**	**4.0**	**-**	**-**	**45.0**	**3.0**	**-**	**-**	**-**	**52.0**	**6.7%**
T9.1	Conceptual map	CWI									12.0					12.0	1.5%
T9.2	Query language and support over language models	CWI									18.0					18.0	2.3%
T9.3	Language models	EXALEAD						4.0			6.0					10.0	1.3%
T9.4	Correlation relevance and derivation of new indicators	FUNREDES									9.0	3.0				12.0	1.5%
																-	0.0%
																-	0.0%
WP 10	**System Development and Integration**	**FRAUNHOFER**	**-**	**5.0**	**22.8**	**7.8**	**5.8**	**2.0**	**3.8**	**45.3**	**3.8**	**-**	**3.8**	**-**	**3.3**	**103.4**	**13.3%**
T10.1	System requirements	FRAUNHOFER			3.0	3.0	2.0	1.0	1.0	6.0	1.0		1.0		1.0	19.0	2.5%
T10.2	Terminology definition	MAAYA		3.0	0.3	0.3	0.3		0.3	0.3	0.3		0.3		0.3	5.4	0.7%
T10.3	System architecture	FRAUNHOFER			3.0	3.0	2.0		1.0	12.0	1.0		1.0		1.0	24.0	3.1%
T10.4	Project infrastructure, process and tools	FRAUNHOFER			15.0					15.0						30.0	3.9%
T10.5	Prototype development cycle	FRAUNHOFER		2.0	1.5	1.5	1.5	1.0	1.5	12.0	1.5		1.5		1.0	25.0	3.2%
																-	0.0%
WP 11	**Applications**	**NIELSEN**	**-**	**20.0**	**-**	**1.0**	**4.0**	**12.0**	**-**	**6.0**	**-**	**-**	**6.0**	**2.0**	**22.0**	**73.0**	**9.4%**
T11.1	Applications evaluation and future impacts forecast	MAAYA		5.0		1.0							2.0			8.0	1.0%
T11.2	Pilot for intelligent information management	EXALEAD					2.0	12.0					2.0			16.0	2.1%
T11.3	Pilot for linguistic diversity public policy	MAAYA		10.0										2.0		12.0	1.5%
T11.4	Pilot for linguistic diversity business	NIELSEN		5.0			2.0			6.0			2.0		22.0	37.0	4.8%
WP12	**Results Assessment & Evaluation**	**FUNREDES**	**-**	**14.0**	**-**	**3.0**	**-**	**-**	**-**	**3.0**	**-**	**18.0**	**-**	**-**	**-**	**38.0**	**4.9%**
T12.1	Internal validity: results crosschecking and comparisons	FUNREDES		2.0								8.0				10.0	1.3%
T12.2	Recognition algorithm method	MAAYA		7.0												7.0	0.9%
T12.3	Words sampling method	FUNREDES		1.0								6.0				7.0	0.9%
T12.4	Utility and Impact assessment	MAAYA		4.0								4.0				8.0	1.0%
T12.5	System testing and validation	FRAUNHOFER								3.0						3.0	0.4%
T12.6	External validity – methodological soundness	DIALOGIC				3.0										3.0	0.4%

Figure 13 - Summary of effort at Task level

B 1.3.4 Risk analysis

B 1.3.4.1 Technology-related risks overview

Risk	Description	Contingency plan or mitigation strategy
R1.1	Poor compression performance. The DILINET platform and the success of the multiple analyses conducted over the sampled data depend on the reliability, efficiency, and scalability of the large-scale repository. Moreover, the data storage will take full advantage of the recent advances in compression algorithms in order to successfully fulfil its efficiency and scalability goals. In the DILINET case, the samples stored in the repository will maximize diversity, and such diversity might degrade the performance of the compression techniques that perform best when self-similarities are high.	Very efficient clustering techniques will be exploited to organize the samples stored in the large-scale repository in buckets maximizing the inter-bucket compression rate. The compression performance will reduce the cost of state-of-the-art solutions granting the reliability and fault tolerance of the data storage component.
R1.2	Poor coverage of crawls. The DILINET crawling strategies relies on research results derived from the study of novel problems. Given the very highly variable nature of the Internet ecosystem we cannot know in advance if our proposals will remain valid also in the near future. Furthermore, since the most interesting content for us is that which is also poorly connected to the mainstream web we may incur in issues regarding the possibility to reach that content.	The risk is mitigated by two factors. Firstly, in DILINET we may also exploit strategies based on other data gathering means. In this case a loss in effectiveness of the crawling phase might be mitigated by the data obtained from the other gathering strategies. Secondly, sampling mechanisms could be adapted to overcome the limitation that crawling might incur in.
R1.3	Concept annotations and/or opinion detection not ready (necessary to integrate in concept maps for D9.6)	Resort to more syntactic annotations for the second user study
	Interactive tool remains too abstract to use in daily life analyst	Researchers create task- and context- specific expressions for the analyst's needs on the basis of interviews and observation study of analyst at work.
R1.4	unable to locate sufficient appropriate audio data	explore alternative sampling schemes with WP5
R1.5	unable to obtain a consensus on what annotations (language e / dialect / content /opinion) are needed for all applications	define a common subset of annotations and extensions for particular uses
R1.6	unable to obtain sufficient annotations with crowd sourcing	contact language schools or other associations to locate annotators
R1.7	automatic speech recognition is insufficient for content detection.	a small amount of data will be manually corrected for a subset of the most important languages for the project
R1.8	Interactive tool remains too abstract to use in daily life analyst	Researchers create task- and context- specific expressions for the analyst's needs on the basis of interviews and observation study of analyst at work.

Risk	Description	Contingency plan or mitigation strategy
R1.9	User accepting plug-in software does not constitute a large enough panel	International Organizations involved in the project will contribute to promotion.
R1.10	Origin of users accepting plug-in software is not varied enough and provokes statistical bias.	Statistical methods will be used to correct the biases.

B 1.3.4.2 Non-technology-related risks overview

Risk	Description	Contingency plan or mitigation strategy
R2.1	Partnership risks: a partner leaves the consortium. Commercialization at risk: industrial partner leaves the market.	In that case, we will promptly identify and recruit a new partner. A specific plan will be defined to bring the new partner up to speed in the shortest time as possible, in order to not cause further delays. In that case, it is possible that a re-allocation of some tasks within the remaining consortium occurs.
R2.2	Defaulting partner: a partner does not deliver his contribution to the work plan, creating severe gaps and delays.	If this occurs, it is likely to be quickly detected by the SCO (task T2.2, Project Technical Assessment that includes the assessment of individual partners). If no corrective action is undertaken by the defaulting partner, the GA has the possibility to suspend payments and to exclude the partner.
R2.3	Management risk: insufficient management activity. Inadequate communication between partners.	Additional resources from ERCIM/EXALEAD will be devoted to project management. Corrective actions will be defined and implemented with the control of the PEB.
R2.4	Conflict within consortium : There is a severe conflict between 2 or more partners of the DILINET consortium that prevents any constructive collaboration.	There is an escalation procedure foreseen in the Consortium Agreement for addressing these situations. Successively, the PEB and the GA may intervene. If no long term solution can be found, then the consortium will take the necessary measures, similarly than in the case of defaulting partners.

Additional risks may be elicited, such as:

- Market risks: nobody is interested by the DILINET platform.
- Regulation/safety risks: new laws and regulations may limit the exploitation potential of the technology.

These risks relate to situations that may occur after the project ends. They will be appropriately managed in due time.

B 2. Implementation

B 2.1 Management structure and procedures

B 2.1.1 Management structure

Project management approach

The DILINET management structure is designed to drive the efficient implementation of the Project's activities as defined in the work plan and the achievement of its scientific objectives, as well as the completion of its contractual obligations vis-à-vis the European Commission, in compliance with detailed rules and procedures.

We recognize that for Projects involving geographically separated participants, a major risk for failure is the lack of coordination. Accordingly, our Project management approach consists of a comprehensive Project management plan together with clear reporting procedures to ensure efficient Project quality and cost control, as well as visibility for all Project partners and external contacts throughout the lifetime of the Project. Special attention is brought into linking all Project components and maintaining smooth communications between the partners.

A management structure that assures close control has been established and well-defined objectives have been set for all partners to ensure agreement even before the Project begins.
Project management in DILINET relies on three principles:

- The creation of a simple and effective management structure to allow for quick decisions.
- The establishment of simple mechanisms to efficiently resolve problems and potential conflicts.
- Responsibilities are clearly defined for self-contained subsets of work to minimize overall risks.

Overall structure

The following diagram presents the overall structure of the DILINET management and reporting structure.

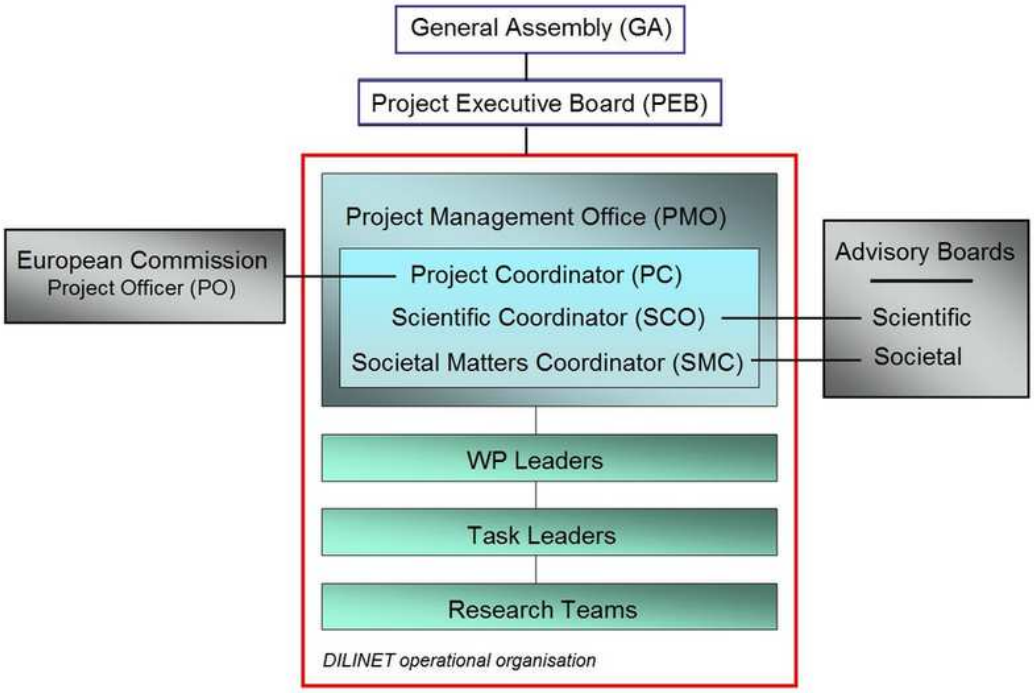


Figure 14 - Management structure

The boxes within the red box consist of the DILINET operational organisation.
The responsibility of each body or role in the organisation is defined in the following paragraphs.

General Assembly (GA)

The General Assembly is the ultimate decision-making body of the Consortium.
The General Assembly consists of one representative (GA Members) of each Consortium partner. The GA is chaired by the Project Coordinator, who initiates an ordinary meeting at least once a year and an extraordinary meeting at any time upon request of the Project Executive Board or of 1/3rd of the Members of the General Assembly.
Each GA Member present or represented in the meeting shall have one vote. Decisions shall be taken by a majority of two-thirds (2/3) of the votes. A decision is being escaladed at the GA level (in administrative, financial, scientific and technical domains) whenever the decision impacts the Project objectives or when the Project Executive Board cannot reach a consensus.

The following decisions shall be taken by the GA only:

- Evolution of the Consortium Agreement, premature termination of the Project.
- Evolution of the consortium (entry of a new partner, termination of a partner's participation).
- Evolution of the Project that impacts the grant with the commission or the content of the Project as defined in the submission documents.
- All budget-related matters, content, finances and intellectual property rights.

The GA decisions are binding to all partners in all Project-related matters. Recommendations provided by the Advisory Boards, representatives of the European Commission, and other Project-related panels will be considered within the decision-making process.

Project Executive Board (PEB)

The Project Executive Board is the supervisory body for the execution of the Project which reports to and is accountable to the General Assembly.
The Executive Board shall consist of the Project Coordinator, the Scientific Coordinator, the Societal Matters Coordinator and Work Package Leaders as appointed by the General Assembly (hereinafter

Executive Members). The Scientific Coordinator (SCO) shall chair all meetings of the Project Executive Board. The Societal Matters Coordinator (SMC) will represent the users' interest and vision in the PEB. The SMC's role gives credit to the historical and actual importance of the user side in the DILINET project and represents a set of motivated International organisations seeking to obtain usable and perennial indicators. The SCO will organise ordinary PEB meetings at least quarterly and extraordinary meetings at any time upon request of any Executive Member. PEB meetings can be organised via conference calls.

Given the definition above, the PEB will therefore be composed of the following 10 members:

PEB member	Organisation	WP leader of
Philippe Rohou	ERCIM	WP1 - Project Management
Gregory Grefenstette (Chair)	EXALEAD	WP2 - Scientific Coordination WP8 - Validity & Analysis
Daniel Prado	MAAYA	WP3 - Dissemination & Exploitation WP4 - Societal Issues
Fabrizio Silvestri	CNR	WP5 - Smart Sampling of Large-Scale Data
Robbin te Velde	DIALOGIC	WP6 - User Centered Measurement
Lori Lamel	CNRS	WP7 - Language Indicators in Non-Text Media
Arjen Devries	CWI	WP9 - Data & Knowledge Representation
Gerhard Paass	FRAUNHOFER	WP10 - System Development and Integration
René Lamsfuss	NIELSEN	WP11 - Applications
Daniel Pimienta	FUNREDES	WP12 - Results Assessment & Evaluation

Figure 15 - Project Executive board

The PEB shall:

- Manage and monitor the effective and efficient implementation of the Project according to the decisions of the GA and prepare the meetings of the GA when new decisions are required.
- Initiate, coordinate and have organised the Work Packages according to the Project work plan.
- Support the SCO in the scientific and technical validation of Project results and external deliverables.
- Support the PC in preparing meetings with the European Commission and in preparing related data and deliverables.
- Support the SCO in preparing meetings with the Advisory Boards.
- Make the decisions related to minor changes in the Work Packages and in the technical roadmap of DILINET (including restructuring Work Packages).

The PEB shall seek a consensus among the partners.

Project Management Office (PMO)

The Project Management Office consists of a team grouping the Project Coordinator (PC), the Scientific Coordinator (SCO), the Societal Matters Coordinator (SMCO) and the chair persons of the two External Advisory Boards (EAB). A detailed definition of EAB's is given below. Philippe Rohou (ERCIM) will act as the Project Coordinator; Gregory Grefenstette (EXALEAD) will act as the Scientific Coordinator; Daniel Prado (MAAYA) will act as the Societal Matters Coordinator. Together they have an extensive qualification, expertise and experience relevant to Project management.

The Project Coordinator shall be the intermediary between the Partners and the European Commission and shall perform all tasks assigned to it as described in the EC Grant Agreement and in the Consortium Agreement. The Project Coordinator is in charge of the administrative and financial management of the Project. In particular, the PC shall be responsible for:

- monitoring compliance by the Partners with their obligations;

- setting-up a collaborative work environment accessible to all Partners;
- keeping the address list of Members and other contact persons updated and available;
- collecting, organising a quality review process and submitting reports and other deliverables (including financial statements and related certifications) to the European Commission;
- transmitting documents and information connected with the Project to and between work package Leaders, as appropriate, and any other Partner concerned;
- chairing the DILINET General Assembly;
- organising the Project review meetings in coordination with the EC Project Officer;
- administering the Community financial contribution and executing the payments to Partners;
- providing, upon request, the Partners with official copies or originals of documents which are in the sole possession of the Coordinator when such copies or originals are necessary for the Parties.

The Scientific Coordinator (SCO) is in charge of the planning, management and monitoring of the research and technological development activities, including the coordination of scientific and technical work between work packages. In particular, the SCO shall be responsible for:

- monitoring the Project progress on a day-to-day basis for continuous rating of the achievements, objectives, tasks, work packages, and the entire Project;
- assessing the scientific contribution of each individual Project Partner;
- ensuring a smooth and efficient collaboration of all Partners;
- chairing the DILINET Project Executive Board;
- keeping close contact with the chair persons of the Advisory Boards;
- leading the scientific dissemination activities;
- checking the delivery of documents and information regarding the DILINET Project within the agreed time;
- driving the process for updating the description of work according to Project, science and technology evolution.

The Societal Matters Coordinator (SMC) is in charge of planning, management and monitoring of the products and activities related to the usage and processing of the research products, including the coordination of work between work packages related to the usage and products of the research (WP3, WP4, WP11, and WP12). In particular, the SMC shall be responsible for:

- monitoring the definition and specifications of the type of data to be produced by the Work Packages dedicated to the research progress;
- assessing the quality of the data produced by the scientific Work Packages dedicated to the research progress and their compliance with the requirements;
- negotiating with the two other coordinators changes required on the definition of the data to be produced;
- ensuring a smooth and efficient collaboration between the research Partners and the usage Partners;
- representing the interests of the usage Partners, and in particular the International Organisations involved in the project (directly as partners or indirectly as sub-contracts);
- keeping close contact with the chair person of the Societal Advisory Boards;
- leading the dissemination and exploitation activities;
- checking the delivery of documents and information regarding the DILINET Project within the agreed definitions;
- propose balanced trade-offs in the process of updating the description of work according to the evolution of project, science and technology on one hand, and the data requirements on the other hand.

Work package Leaders are responsible for the coordination and monitoring of all the activities composing in their work package and for the liaison with the Project Management Office and the other

work package leaders. They will organise work package meetings as required, using extensively dedicated tools ranking from mailing lists to audio-conferencing services. They will promptly report to the PMO any deviation with respect to the Project plan in order to implement fast corrective actions.

Task Leaders are responsible for the coordination and monitoring of all the activities composing their task and for the liaison with the work package leader and the other task leaders. They will organise task meetings as required, using extensively dedicated tools ranking from mailing lists to audio-conferencing services. They will promptly report to the work package leader any deviation to the Project plan.

External Advisory Boards (EAB)

Considering that the current project has the potential to profoundly impact society and/or industry it is important that large cultural players and opinion leaders are involved. Therefore two advisory boards will be installed containing representatives of companies as well as institutions:

- The Scientific Advisory Board will be chaired by Prof. Ricardo Baeza Yates (VP of Yahoo Research, Europe),
- The Societal Advisory Board will be chaired by Mrs Vanessa Gray (Senior ICT Analyst, ICT Data & Statistics in ITU)

So far, the following people have accepted to join the DILINET advisory boards:

Scientific Advisory Board:
Mikami Yoshiki, Leader of Language Observatory Project, Professor Nagaoka University of Technology
Joseph Mariani, CNRS

Societal Advisory Board:
Alexandre Wolff, OIF, Responsable de l'Observatoire de la langue française
Fabio Nascimbeni, Diector, Menon Network
Mikami Yoshiki, Leader of Language Observatory Project, Professor Nagaoka University of Technology
Mrs Dolores Alvarez, Union Latine
Académie de l'intelligence économique

Mrs. Gray will liaise with the members of the Partnership on Measuring ICT for Development and invite some members to join, such as Eurostat, OCDE, UNESCO/UIS, UNCTAD, UNDESA.

These Advisory Boards will have three specific missions:

- Monitor project progress on achieving major objectives.
- Foster innovation: preparing the adoption of DILINET resulting technologies by the different stakeholders, by providing input regarding the required outcome parameters, the usability, application scenarios and opportunities.
- Legal and ethical guidance: providing guidance to DILINET research groups to ensure that activities and technologies (carried out during the project or considered for the further exploitation of project results) are respecting European legislation and ethical principles.
- Upon request, evaluate and assess strategic reports produced by the DILINET work packages.

The Advisory Boards will meet (1-day meetings) once a year. More frequent interaction (by email and/or audio conference) may be set up, depending on project needs and on Advisory Board members' interest.
A "Memorandum of Understanding" will be put in place, defining the rights, duties and expectations between the Advisory Board members and the DILINET consortium.

Dissemination and Exploitation Manager

Since exploitation and dissemination tasks are essential for the success, usage and acceptance of the projects outcomes, a Dissemination and Exploitation Manager (DEM) will be in charge of the coordination of such tasks in the project. This will be the role of the WP3 leader. Concerning dissemination he will be responsible of all the issues related to the wide diffusion of the project results in order to ensure the largest possible visibility and successful exploitation.

At the meetings of the Project Executive Board, the DEM will introduce the dissemination agenda and address all the issues related to its attributions (publication, participation to professional and public events, awareness raising activity). Concerning exploitation, he will be in charge of liaison with all other partners relating to exploitation, negotiation with external partners, and IPR (patents, licensing, and royalties).

Quality management

Quality management is of the highest importance for maximising the chances to reach all Project objectives. Quality management is driven by the Project Management Office. On the administrative level, the PC will set up and organise the quality assurance process for all Project deliverables. On the scientific and technological level, the SCO will assess the quality of the contribution of all partners and of Project results.

Intellectual Property Rights (IPR) management

Proper IPR management aims at setting the basis for a successful exploitation of the Project results. The general rules and procedures related to IP are defined in the Project Consortium Agreement. Daniel Pimienta (FUNREDES) will act as IPR manager that will be in charge of driving the IPR-related processes, and possibly setting up an IPR Committee to resolve specific issues.

B 2.1.2 Procedures and tools

The following procedures and tools will be used in order to ensure an efficient management and communication throughout the whole Project organisation.

Management notes

Project specific processes (e.g. deliverable review, publication approval) will be documented in Management Notes that will be produced by the PMO. Management notes will be written by the PC and approved by the SCO before being sent to the Partners.

Collaborative Tools

The PMO will set up a collaborative working environment, consisting of following items:

- a set of Mailing Lists, to ensure an efficient communication (including archiving of messages) between the different communities of the project (GA, PEB, PMO, AB, WP Leaders, etc.).
- a Multilingual platform allowing i) discussions between partners with the possibility for each participant to see contributions in his/her own language and ii) a feature allowing the sharing of working documents.
- a Collaborative web environment (BSCW or equivalent) to be a repository for all information generated by the Project, such as contractual documents, project deliverables, minutes of meetings, dissemination material, etc.

Meetings

It is expected that the Project Executive Board will organise a monthly audio call and will meet face-to-face every quarter (one of which to prepare and attend the annual EC review). The Advisory Boards will meet face-to-face once a year, in parallel with one of the PEB meetings. It is assumed that because all project partners will (be invited to) attend all face-to-face PEB meetings, any of the quarterly PEB meeting can designated as a GA meeting.

B 2.1.3 Conflict resolution

Even if the preferred decision making process is aimed at building consensus between the partners, a divergence or a conflict between several parties may arise. The following topics may cause such conflict:

- Technical discussion unresolved,
- Task allocation,
- Partner not delivering,
- conflict between persons,
- Funding distribution,
- IPR,
- Etc.

The DILINET overall structure provides a clear way to manage conflicts, and with respect to the escalation of conflict resolution (up to GA), the Consortium Agreement describes the process for settlement of disputes. The escalation process is defined as follows:

- Search for a solution by the Project Management Office,
- If no resolution is found, then involve the Project Executive Board, which is empowered to decide on minor issues,
- If no resolution is found, then involve the General Assembly, which is empowered to decide on major issues,
- Whenever appropriate, maintain a close communication with the EC Project Officer.

In order to ensure an efficient operational management of the Project activities, it is very important that the PMO expresses a unique point of view: the PC, the SCO and the SMC will therefore maintain a very close communication channel and seek consensus between them. The risk of conflict between the PC and the other two coordinators is minimal because (i) the role of each one is clearly defined and (ii) the nature of their organisation makes it unlikely to have a conflict of interest. A sane and fruitful tension may arise from the existence of both a SCO and a SMC. This healthy tension will result in a pressure on researchers to deliver the data closest to what users expect on one hand, and will produce a better understanding from the users of the difference between the field of possible and the field of wishes, on another hand. Because either coordinator has sufficient knowledge of the remit of the other (research vs. indicators), we anticipate a smooth and creative decision process concerning the requirements. If ever required, the PC will play a role of arbitration.

B 2.2 Beneficiaries

Partner 1: GEIE ERCIM, France (Coordinator)

The **European Research Consortium for Informatics and Mathematics** (ERCIM, www.ercim.eu) is a Consortium of two organisations, a European Economic Interest Grouping (EEIG), and a Non Profit International Association (AISBL), composed of a network of research institutes from twenty two European countries, embodying more than 12,000 researchers and engineers. ERCIM is based in Sophia Antipolis (France) with an antenna in Brussels.

ERCIM's mission is to: foster collaborative work within the European research community in Information and Communication Technologies (ICT) and Applied Mathematics; advise the European Commission and national governments; and increase co-operation with European industry. ERCIM is also the European host of the World Wide Web Consortium (W3C), whose mission is to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the web.

Role in DILINET
ERCIM will lead WP1 and provide the financial and administrative coordination of the project and will contribute to the dissemination activities (in particular via the ERCIM News and ERCIM Innovation publications). ERCIM will also call upon its W3C staff for punctual support with specific issues related to language standardisation and normalisation.

Expected outcome from DILINET
Through DILINET, ERCIM is fully accomplishing its mission, supporting the European leading academies and industries in their respective search for scientific and business excellence.

Key personnel

Philippe Rohou will serve as the Project Coordinator of DILINET, and will personally lead all tasks under ERCIM responsibility. His European project management experience includes notably the administrative and financial coordination of the DELOS NoE (60 partners), of the CoreGRID NoE (42 partners), of the RACE-network RFID thematic network (25 partners), of the D4Science I3, of the Digital World Forum CSA, of the AXES IP, of the Net-WMS STREP, and a few others, as well as dissemination work package leadership for the VPH NoE.

Partner 2: Réseau Mondial pour la Diversité Linguistique, CH (MAAYA)

An initiative that came out of the second phase of the World Summit on the Information Society (WSIS) in Tunis in November 2005, the World Network for Linguistic Diversity, MAAYA (http://maaya.org), aims to value linguistic diversity as a building block of unicity of human communications. MAAYA serves as a multi-stakeholder network in the area of shared knowledge, where technology offers a great potential for languages, but is also a risk to them. MAAYA is as a focal point for linguistic research projects and its objectives includes the promotion of software localization and equal access of all languages to cyberspace and the observation of the implementation of language policies. The founding members of MAAYA are : African Academy of Languages (ACALAN), Linguamón - Casa de les Llengües, Codice Idee per cultura SRL, E-Africa Commission du NEPAD, ENSTA, Funredes, Global Knowledge Partnership , GREF, SIL International, Linguasphere Observatory,Intlnet, ICVolunteers, Institut francophone des nouvelles technologies de l'information et de la formation (INTIF), Language Observatory, International Literacy Institute of the University of Pennsylvania (ILI), Multilingual Internet Names Consortium (MINC), Organisation Intergouvernementale de la Francophonie (OIF), RECLA, Thai Computational Linguisitcs Laboratory, Toile Métisse, UNESCO, Unicode IDN in Africa, African Union, International Telecommunication Union (ITU), Union latine.

Role in DILINET
MAAYA is the main partner in terms of use of the research results, coordinating the implication of some of its members and the international organizations involved in DILINET. It has the responsibility of two work-packages dealing with societal considerations raised by the DILINET research and their dissemination and exploitation and is assuming the role of coordination for Societal Matters.

Expected outcome from DILINET
The expected outcomes of DILINET in terms of creating indicators and demonstrating the usefulness of linguistic diversity measurement through public policies and economical impacts are totally coherent with the international mission of MAAYA and pave the ground towards the World Summit on Multilingualism.

Key personnel

Daniel Prado is the Executive Secretary of MAAYA. From 1984 to 2011, he was responsible of the Directorate for Terminology and Language Industries (DTIL) of the intergovernmental organisation Union Latine
Within the framework of the DTIL, he has promoted the modernization of the romance languages in order to facilitate access to specialized communication in mother tongue. Daniel Prado has coordinated activities related to development of terminologies and language tools, the promotion of the scientific and technical translation and writing, and also the development of linguistic diversity in cyberspace, science, international negotiations, international governance and more. He has been responsible for numerous projects involving national and international spheres in language policy, terminology, language industries and multilingualism in cyberspace, as well as building information systems and specialized multilingual sites (www.portalingua.info, www.terminometro.info, www.hex-libris.info, etc.) and fora. He participated in the creation of several international networks or associations such as Realiter, RITem, Linmiter or EAfT. He coordinates the reference multi-author book "Net.Lang-Challenges of multilingualism in cyberspace", the preparation of the Third International Symposium on Multilingualism in Cyberspace (Paris, November 2012) and activities for the realization of the World Summit on multilingualism.

Publications

- “Diversité linguistique et cyberespace : état de l'art, enjeux et opportunités”, D. Prado, D. Pimienta, A. Lemoulinier, Cosmopolis, 2010
- "La traduction automatisée: le cas des langues romanes" in Revue Hermès Nº 56 Traduction et mondialisation, 2010
- "Diversité linguistique et cyberespace : état de l'art, enjeux et opportunités"[40] - Daniel Prado, Daniel Pimienta, Anneflore Lemoulinier in Cosmopolis (01/10)
- "Languages and cyberespace" in International Conference on Linguistic and Cultural Diversity in Cyberspace Proceedings (Yakutsk, Russian Federation, July 2-4, 2008)

Partner 3: Istituto di Scienza e Tecnologie dell’Informazione, IT (CNR)

The Information Science and Technologies Institute (ISTI) (http://www.isti.cnr.it/) is an institute of the Italian National Research Council (CNR) (http://www.cnr.it), located in Pisa. ISTI is actively involved in collaborations with the academic world and in cooperative research and development programmes, both national and international. The Institute is committed to producing scientific excellence and to playing an active role in technology transfer. The domain of competence covers Information Science, related technologies and a wide range of applications. The HPC Lab (@ISTI) conducts research in various areas various areas of High Performance Computing:

Cloud and P2P Systems;
Data and Web Mining;
Information Retrieval;
Distributed Search Engines.

The Laboratory’s research funding comes through national, EC, and industry projects which span many areas. The group, composed of about 20 people, is actively involved in several important European projects including S-Cube, Contrail, Assets, and XtreemOS in the seventh framework programme. The specific HPC Lab experience, relevant to the project, is in the area of efficiency issues of IR systems, indexing and techniques for improving query processing performance. These technologies are studied mostly in the context of distributed search systems.

[40] http://agora.qc.ca/cosmopolis.nsf/Articles/no2010_1_Diversite_linguistique_et_cyberespace___etat_de_l?OpenDocument

Key personnel

Fabrizio Silvestri is currently a Researcher at ISTI - CNR in Pisa. He received his Ph.D. from the Computer Science Department of the University of Pisa in 2004. His research interests are mainly focused on Web Information Retrieval with particular focus on Efficiency related problems like Caching, Collection Partitioning, and Distributed IR in general. In his professional activities Fabrizio Silvestri is member of the program committee of many of the most important conferences in IR as well as organizer and, member of the steering committee, of the workshop Large Scale and Distributed Systems for Information Retrieval (LSDS-IR). He is author of more than 80 publications in highly relevant venues spanning from distributed and parallel computing to IR and data mining related conferences. He has been the recipient (together with Ranieri Baraglia) of the best Web Intelligence 2004 paper award, and the recipient of the ECIR 06 best paper award. Recently Fabrizio Silvestri has been appointed as Work-Package leader and Activity leader of the EU Network of Excellence project S-Cube. In addition, he serves as a member of the steering committee of S-Cube Project. Fabrizio Silvestri has written recently a survey paper for the journal Foundations and Trends in Information Retrieval, and has given a keynote speech at the LA-Web 2008 conference with a talk entitled ìPast Searches Teach Everything: Including the Future. He is the Program Commitee Chair for the IR track of SPIRE 2011 in Pisa. He is the IR Track Program Committee Chair of the AICCSA 2011 conference. He is the Area Chair for “Search Engine Architectures and Scalability” at SIGIR 2011. He is the Industry Track co-chair at ECIR 2011. He is senior PC member of the CIKM 2011 conference.

Claudio Lucchese graduated in Computer Science at the University "Ca' Foscari" of Venice, summa cum laude on february 15th 2002. At the same university he got the Master Degree in Computer Science summa cum laude, on October 10th 2003. The title of his theses is "Algorithms for the extraction of frequent itemset and closed itemset". During 2004, he has been a research associate at the Information Science and Technology Institute (I.S.T.I.), which belongs to the italian National Research Council (C.N.R.), where he worked for the High Performance Computing Lab. (H.P.C.) and collaborated with the Knowledge Discovery and Delivery Lab. (K.D.D.) in the project P3D: Privacy Preserving Pattern Discovery. He received the Ph.D. in Computer Science on February 18th at the the "Ca’ Foscari" University of Venice under the supervision of prof. S.Orlando [web]. The title of the Ph.D. thesis is "High Performance Closed Frequent Itemsets Mining inspired by Emerging Computer Architectures". Since November 2007, Claudio Lucchese is a researcher at the I.S.T.I.-C.N.R. in Pisa. This turned into a permanent position on June 2008. He is author of more than 50 publications on Data Mining and High Performance Data Management. He has been actively involved in the organisation of a very popular workshop on Large Scale and Distributed Systems for Information Retrieval (LSDS-IR) that is usually held in conjunction with the most important conferences in Information Retrieval.

Partner 4: DIALOGIC, NL (DIALOGIC)

Dialogic is an independent research-based consultancy firm located in Utrecht, the Netherlands. Dialogic focuses on processes of innovation IT, telecom, new media developments, and services. Dialogic employees have been active in these fields for a long time and have performed many projects in the area of telecommunications and media studies, conditions for introducing new technologies, assessments of user requirements, technology-based foresight studies, scenario-studies and economic industry analysis. In the area of telecommunications and media, Dialogic has developed towards an established centre of expertise, both covering the Netherlands and increasingly Europe. In particular, comparative and benchmarking studies have been performed, in the areas of broadband internet penetration and user pattern development, and the development of user needs. With the internet as data source project, Dialogic has introduced the automated collection of data from the internet for statistical purposes (including the notion of user-centric measurements). The initial project (2006/2007)

was geared towards the Dutch Bureau of Statistics. Follow-up projects have recently been completed (EC) or are planned for 2012 (Eurostat, OECD, IPTS).

Role in DILINET

Dialogic will offer its general expertise on internet-based data collection models and will develop and built the user-centric measurement client for language detection in actual use. Dialogic will also provide parts of the overall quality management of the project (audit external validity of the research results).

Expected outcome from DILINET

To application of user-centric measurements at this level of detail is a new and unique development track. Moreover the combination of the client with a survey model that can be triggered by the data from the client renders it into a very powerful panel survey instrument. The language component is entirely new for Dialogic and is an exciting new area to enter.

Key Personnel

Robbin te Velde is principal researcher at Dialogic. He has extensive experience in ICT research projects. During the last 20 years he has been alternatively working at technical universities (Twente University, Delft University of Technology, Eindhoven University of Technology, research consultancies, and think tanks (including the Rand Corporation). He has a strong background in methodology and is specialized in international comparative studies. Next to strategic IT consulting he has implemented hands-on IT-projects for large telecom operators (including BT and KPN) and several multinationals. Besides completing over 100 research projects he has written a large number of scientific articles on a wide range of areas such as international politics, philosophy, knowledge management, business administration, technology policy and information management. Robbin has been principal researcher in all Dialogic projects with regard to automated data collection.

Reg Brennenraedts is senior researcher and consultant at Dialogic. He holds a bachelor in Electrical Engineering, a Master of Science in Innovation Science (Eindhoven University of Technology) and an MBA in strategy and corporate finance (TiasNimbas Business School). He has been mainly working in the fields of telecom and IT. He has a regular advisor to the Dutch telecom operator and Dutch Ministry of Economic Affairs and has been involved in various broadband projects throughout the Netherlands. With regard to IT he has been involved in innovation projects with regard to gaming, digital music, digital radio (DAB) and digital TV. Reg is a seasoned project leader and has also been in charge of all automated data collection projects at Dialogic.

Partner 5: Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, FR (CNRS)

Founded in 1939, the CNRS (National Centre for Scientific Research) is a government-funded research organisation, under the administrative authority of France's Ministry of Research (http://www.cnrs.fr). The CNRS encourages collaboration between specialists from different disciplines and has laboratories located throughout France. LIMSI is one of France's largest research laboratory working on language technologies; it covers the full spectrum from low level signal processing to spoken and written language processing and machine translation. The main activities of the Spoken Language Processing Group cover the following domains: Speech Recognition, Speech Understanding, Dialog Systems, Speaker and Language Recognition, Speech Translation and Audio Indexation. Associated activities include data collection, system evaluation and technology transfer. The group has succeeded in basic research as well as in applied research developing new algorithms, prototypes and databases. Advanced commercialized products developed from studies at LIMSI are now being used in several applications. LIMSI has participated in a number of projects on speech recognition (CORETEX), spoken language

systems for information retrieval (EC MASK, Railtel, Arise, Home, DISC and Amities) and audio document indexation and retrieval in multiple languages (EC Olive, Alert, Echo, RNRT Theoreme, AudioSurf), facilitating human-human communication (CHIL) and spoken machine translation (TC-STAR). LIMSI, in collaboration with Vocapia Research, has developed competitive transcription systems for broadcast data in 9 languages in the context of the Quaero program. (Web site: http://www.limsi.fr/tlp)

Role in DILINET

The main contributions from CNRS-LIMSI to DILINET will be in the development of robust statistical models for language identification, multilingual speech recognition, content and opinion extraction for the wide variety of audio data types found on the web.

Expected outcome from DILINET

The CNRS will improve their models and technology for language recognition and content extraction in heterogeneous data for a large number of languages. DILINET will also provide a means to better access to language resources in many languages.

Key personnel

Lori Lamel is a senior CNRS researcher in the Spoken Language Processing group at LIMSI which she joined in October 1991 She received her PhD degree in EECS in May 1988 from the Massachusetts Institute of Technology. Her principal research activities are in speech recognition; acoustic-phonetic studies; lexical and phonological modeling; and conversational systems. She has been a prime contributor to the LIMSI participations in speech recognizer benchmark evaluations and developed the American English pronunciation lexicon. She has been involved in many European projects, most recently leading the LIMSI activities in the IP Chil. Dr. Lamel is a member of the Speech Communication Editorial Board, was a member of the Interspeech International Advisory Council, the IEEE James L. Flanagan Speech and Audio Processing Award Committee (2006-2009) and the EU-NSF Working Group for 'Spoken-Word Digital Audio Collections'. She has over 230 reviewed publications and is co-recipient of the 2004 ISCA Best Paper Award for a paper in the Speech Communication Journal.

Jean-Luc Gauvain is a senior researcher at the CNRS, where he is head of the Spoken Language Processing Group at LIMSI. He received a doctorate in Electronics from the University of Paris-Sud 11 in 1982, and joined the CNRS as a permanent researcher in 1983. His primary research centres on large vocabulary continuous speech recognition and
audio indexing. His research interests also include conversational interfaces, speaker identification, language identification, and speech translation. He has participated in many speech related projects both at the French National and European levels and has led the LIMSI participation in DARPA/NIST organized evaluations since 1992, most recently for the transcription of broadcast news data and of conversational speech. He has over 250 publications and received the 1996 IEEE SPS Best Paper Award in Speech Processing and the 2004 ISCA Best Paper Award for a paper in the Speech Communication Journal. He was co editor-in-chief of the Speech Communication from 2007 to 2009.

Publications

L. Lamel and J.L. Gauvain. Speech recognition. In R. Mitkov, editor, OUP Handbook on Computational Linguistics, chapter 16, pages 305322. Oxford University Press, 2003.

J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. Speech Communication, 37(1-2):89108, 2002.

L. Lamel and J.L. Gauvain, A Phone-based Approach to Non-Linguistic Speech Feature Identification, Computer Speech and Language, 9(1):87-103, January 1995.

M. Faouzi BenZeghiba, J.L. Gauvain and L. Lamel, Gaussian Backend Design for Open-set Language Detection, IEEE ICASSP'09, Taipei, April 2009

M. Faouzi BenZeghiba, J.L. Gauvain and L. Lamel, Language Score Calibration using Adapted Gaussian Back-end, Interspeech'09, Brighton UK, September 2009

Partner 6: EXALEAD, FR (EXALEAD)

Founded in 2000 by search engine pioneers, Exalead is a global software provider in the enterprise and Web search markets. Exalead worldwide client base includes leading companies such as Price Waterhouse Cooper, Michelin, American Greetings and Sanofi Pasteur, and more than 100 million unique users a month use Exalead's technology for search. Today, Exalead is reshaping the digital content landscape with a platform that uses advanced semantic technologies to bring structure, meaning and accessibility to previously unused or under-utilized data in the disparate, heterogeneous enterprise information cloud. The system collects data from virtually any source, in any format, and transforms it into structured, pervasive, contextualized building blocks of business information that can be directly searched and queried, or used as the foundation for a new breed of lean, innovative information access applications. Exalead's technology provides users with a single access point to information, regardless of format or location. Its patented Search by Serendipity® navigation system adapts to user habits. Exalead devotes substantial efforts in supporting research and innovation for helping customers succeed and to be in the forefront of emerging technology developments. Therefore, Exalead is engaged in European and French research projects with academic partners and some of the industry's top research organisations to forge new ground in the analysis, classification and usage of digital multimedia content, ranging from text, speech, and music to images and video.

Role in DILINET
Exalead will offer expertise in Web crawling and web page processing (detection of spam, recognition of networks). Exalead will also provide scientific coordination between the research partners.

Expected outcome from DILINET
Exalead recognizes that Big data is the future of the Web, and is eager to be able to exploit its experience in Web crawling to better understand what data is available on the Web. Language models will also be useful in ranking and classification in its search engine.

Key personnel

Gregory GREFENSTETTE, received his PhD from University of Pittsburgh in 1993 and is Chief Science Officer at Exalead. Dr Grefenstette has been a fixture of the information retrieval and natural language processing fields since 1988. He was previously Principal Research Scientist at Clairvoyance Corp (Pittsburgh; 2001-2003) and Principal Scientist at the Xerox Research Centre Europe (Grenoble; 1993-2001), where he managed a team of researchers working on information extraction and information retrieval, and advanced natural language processing applications. He served on more than 30 program committees for conferences on information retrieval, natural language processing and computational linguistics and he is on the editorial board of the Journal of Natural Language Engineering.

Amar-Djalil MEZAOUR, received his PhD from University of Paris Sud (Paris XI), in 2005 and is Research and Development project manager at Exalead. He was previously Researcher Assistant at LRI in IASI team (Artificial Intelligence team) and member of the INRIA GEMO project. Within the research and innovation group of Exalead, Dr Mezaour worked on integrating advanced semantic technologies for information extraction in Exalead software. He also managed European and French research projects

like ALIS (IST FP6 call and WebContent). In Quaero, he is in charge of developing machine translation tools operating on multimedia content (text, audio speech and video).

Partner 7: Universitat Pompeu Fabra, Barcelona, ES (UPF)

The Universitat Pompeu Fabra was founded in 1990 by the Catalan parliament and lies in the very heart of Barcelona. From its beginnings, this young academic institution has set itself two main objectives: to train professionals and citizens who are responsible and committed to civic values, and to contribute towards the development of research. Despite its young age, the Universitat Pompeu Fabra has become one of the top universities in Spain, as witnessed by its first place ranking in scientific productivity in Spain in 2009.

The Web Research Group of the UPF is participating in this project. The Web Research Group is a multi-disciplinary group focussed on the study of Social Media and the Web. It consists of researchers with a track record in Computer Science, Mathematics, Software Engineering, Library Sciences, Natural Language Processing and Computer-Human Interaction. The lab is most known for its contributions in mining the usage, content and structure of the web, text and multimedia information retrieval, semi-structured data retrieval, information visualization and social network analysis. The Web Research Group is closely collaborating with Yahoo! Research Barcelona and the Barcelona Media Foundation.

Role in DILINET

UPF contributes mainly to the information retrieval and structural analysis parts of the project. This includes strategies for crawling and sampling parts of the Web, and network link analysis.

Key personnel

Vladimir Estivill-Castro is currently team leader of MI-PAL, recently classified for RoboCup 2011 in the Standard Platform League using the Humanoid Robot Nao. His main interests are algorithmic engineering, computational complexity, intelligent data analysis, privacy preserving data mining and knowledge discovery. Prof. Estivill-Castro holds a Ph.D. from the University of Waterloo in Canada and degrees from UNAM in Mexico. He serves in the editorial board of the Journal of Research and Practice in Information Technology, is editor in chief of the series Conferences in Research and Practice in Information Technology and serves on the editorial review board of the International Journal of Data Warehousing and Mining. He has been awarded a national citation by the Australian teaching and Learning for his achievements in models of PhD supervision in Computer Science. He has secured over $1M in competitive funding research projects and has over 100 scientific publications.

David F. Nettleton received the BSc (Hons) in Computer Science- from the University of Wolverhampton, U.K., in 1984, the M.Sc. in Computer Software and Systems Design from The University of Newcastle-upon-Tyne, U.K., in 1985, and the PhD degree in Artificial Intelligence from the Polytechnical University of Catalunya, Spain, in 2002. He currently works as Contract Researcher for the Institute for Investigation in Artificial Intelligence-CSIC, Spain, and also with the Web Research Group of the Department of Technology at the Pompeu Fabra University, Barcelona. His current research interests include data privacy and data mining applied to online social net- works represented as graphs. Previously, he has worked for IBM Global Services as a Data Mining specialist and for EDS in software development. He has been a member of the Spanish Computer Technicians Association (ATI) since 1995, is currently board member of ATI-Catalunya, IEEE Member and ACM Professional Member. He has published more than 30 refereed papers in conferences and journals, as well as 2 full books: "commercial data analysis" and "techniques for clinical data analysis".

Anton Dries received two MSc (2004, 2006) and a PhD in Engineering: Computer Science (2010) from the Katholieke Universiteit Leuven, Belgium.
Currently, he is a post-doctoral researcher in the Web Research Group, Technology Department, Universitat Pompeu Fabra. His research interests include general machine learning and data mining with a special focus on data streams and information networks.

Partner 8: IAIS/Fraunhofer, DE (FRAUNHOFER)

The Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (Fraunhofer Society for the Advancement of Applied Research (FRAUNHOFER)) is Germany's leading organisation of research institutions for applied research. Fraunhofer's 60 research institutes are co-ordinating and/or contributing to large national and international industrial applied research projects, as well as to research projects targeting the service sector, national and regional governments and the EU. A staff of 18 000, the majority of whom are scientists and engineers (with university degree), generate the annual research budget of more than €1.6 billion.

The Knowledge Discovery Department (KD) at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in Sankt Augustin, Germany, is a research group (50 scientists) located in the field Machine Learning and Data Mining. Professional experience and expertise in this group include Data Mining, Text Mining, Statistical Relational Learning, Geographic Information Systems, and Distributed Computing. The group has extensive experiences with EU projects, having coordinated in the last 6 years five ICT projects and participated in more than 18. To mention a few, FRAUNHOFER coordinates the LIFT project (local inference in massively distributed systems, 2010-2013), the AntiPhish project (Anticipatory Learning for Reliable Phishing Prevention, 2006-2009) and the KDubiq Coordination Action (Knowledge Discovery in Ubiquitous Environments, 2006-2008), and is involved in the ACGT project (Advancing Clinico-Genomic Trials on Cancer, 2006 - 2009), the LifeWatch project (development of an European infrastructure for coping with the biological diversity, 2007 – 2009), the SCY project (Science Created by You, 2008-2012), the DICODE project (Mastering Data-Intensive Collaboration and Decision Making, 2010-2012), and the IP SIMDAT (Grids for Industrial Product Development, 2004-2008). The group is also currently involved in several nationally funded and industrial projects in Multimedia Mining, Spatial Data Mining, and Text Mining.

Role in DILINET
Fraunhofer will provide expertise in intelligent sampling and multilingual content analysis (detection of content categories and opinions on concepts). Fraunhofer will also coordinate System Development and Integration of DILINET.

Expected outcome from DILINET
Fraunhofer will extend its expertise on sampling and content recognition to a Big Data environment. Especially relevant is the system architecture and system development in this case. We envision many and very relevant political and commercial applications of multi-lingual opinion mining on Big Data.

Key personnel

Gerhard PAASS studied mathematics and computer science and received a PhD in economics at the University of Bonn. Dr. Paaß led the EU project DIASTASIS on text classification for Web mining. He was co-ordinator of the EU project AntiPhish (Anticipatory Learning for Reliable Phishing Prevention, 2006-2009) for filtering phishing emails. He also led several industry projects, e.g. the project MediaRank for the assessment of commercial text classification systems and a project to classify fraudulent eBay offers. He organized a number of workshops on multimedia learning, ontology learning, and text mining for

security at international conferences and served on the program committee for several international conferences and journals like such as ECML-PKDD, UAI, KDD, ICDM, SDM, SIGIR, MLJ, and DAMI. Currently, he leads a subproject of the German joint THESEUS program (New Technologies for the Internet of Services, 2006-2012) on semantic technologies.

Florian SCHULZ majored in computer science at the University of Bonn. He works at Fraunhofer IAIS in a number of projects as programmer, system engineer, and system architect. Most notably are the EU-Project SCY (Science Created by You, 2008-2012) where he designed and implemented an agent architecture capable of automatic and parallel processing of user generated objects. In the DDB (German Digital Library, 2010-2012) he was responsible for designing main parts of the architecture and developing a system integrations process. In this project Florian Schulz also conceptualized and implemented a system to collect digitalised objects from different cultural institutions to store and present them in a unified way.

Michael MAY, obtained his PhD from the graduate program in Cognitive Science, Univ. Hamburg, working on machine learning of causal relationships. He is head of Fraunhofer IAIS Knowledge Discovery department and leads the research efforts at the intersection of data mining, machine learning, and spatial technologies. His current main research interest is application of data mining to structural data. He was and is coordinator of several European projects, including the FET-Open LIFT (Local Inference in Massively Distributed Systems, 2010-2013), FET-Open KDubiq Knowledge Discovery in Ubiquitous Environments Coordination Action (2005-2008) and the KDNet Knowledge Discovery Network of Excellence (2002-2005). He has been local chair of the International Conference on Machine Learning ICML 2005, of ILP 2005, and chairman of the working group Data Management in the FP6 EU Grid Concertation Forum from 2004 to 2006. Dr. May has been principal investigator in several recent industry funded projects on spatial learning.

Partner 9: Stichting Centrum voor Wiskunde en Informatica, NL (CWI)

The Stichting Centrum voor Wiskunde en Informatica (CWI) is the Dutch national research institute for mathematics and computer science. It is a private, non-profit organization located at the Science Park Amsterdam. CWI's mission is twofold: To perform frontier research in mathematics and computer science, and to transfer new knowledge in these fields to society. CWI actively pursues joint projects with external partners, provides consulting services, and stimulates the creation of spinoff companies. CWI also manages the Benelux Office of the W3C and hosts both the Semantic Web Activity Lead and the chair of the XHTML and XForms Working Group. CWI is strongly embedded in Dutch university research: about twenty-five of its permanent senior researchers hold part-time positions as professors at universities and many projects are carried out in cooperation with university research groups. CWI receives a basic funding from the Netherlands Organisation for Scientific Research (NWO), amounting to about two third of the institute's total income. The remaining third is obtained through national research programmes, international programmes, and contract research commissioned by industry. CWI hosts a staff of 235 full time employees, 50 permanent scientific staff, 135 temporary scientific staff, and 50 support staff.

Role in DILINET

Centrum Wiskunde & Informatica brings in their expertise on data management and information retrieval, in particular a long track record of research and applications where the two areas meet. In the VITALAS (FP6 IP) project for example, the XML information retrieval system developed in this group performed the key search operations, including the analysis of user interaction logs for the suggestion of keywords and (multimedia) concepts. The objective of a declarative language for defining and manipulating language models over large amounts of data fits perfectly with the long term focus on the integration of databases and information retrieval. Ongoing research in the Interactive Information

Access group (INS2) addresses more related topics, including user search log analysis, entity retrieval and social media, focusing on how structured information acquired from linked open data can augment interactive information access to semi-structured text and multimedia collections.

Expected outcome from DILINET

CWI will deepen its experience with language modelling approaches, in a much more diverse and heterogeneous setting than the cultural heritage and enterprise data we have worked with before. The methods for inducing topic maps to summarize large amounts of web data will help progress in our research programme where we mix statistical and logical models to further information retrieval. We finally expect to exploit the gained expertise in handling large amounts of multi-lingual web data together with spin-off company Spinque.

Key personnel

Arjen P. de Vries is a tenured researcher at CWI leading the Interactive Information Retrieval research group, and a full professor (0.2 fte) in the area of multimedia data management at the Technical University of Delft. De Vries received his PhD in Computer Science from the University of Twente in 1999, on the integration of (multimedia) information retrieval and database systems. He is especially interested in the design of database systems that support search in multimedia digital libraries. He has worked on a variety of research topics, including (multimedia) information retrieval, database architecture, query processing, retrieval system evaluation, and ambient intelligence. In recent research, he concentrates on the question how to exploit the traces left by people online as links, clicks, and participation in social networks, to improve the quality of information search. He has participated in EU projects VITALAS and PuppyIR, Dutch national programmes MultimediaN and the new COMMIT, and is a member of Cost Action MUMIA and the EU PetaMedia Network of Excellence. He has supervised several best student papers at ACM conferences. He has been general co-chair of the ACM SIGIR 2007 conference in Amsterdam and programme co-chair of CIKM 2011 and ECIR 2012. He coordinates international benchmarking activities at TREC (ranking entities on the WWW). De Vries is a member of the TREC PC, and a steering committee member of INEX (the Initiative for the Evaluation of XML Retrieval).

Jacco van Ossenbruggen is affiliated with the Interactive Information Access group at the Centrum voor Wiskunde en Informatica (CWI) in Amsterdam, and with the Web & media research group at VU University in
Amsterdam. His research interests include user interfaces for unreliable data, web-based metadata modeling and integration, and data provenance on the web. He is currently researching these topics in the cultural heritage domain (as part of the European PrestoPrime and the Dutch national COMMIT projects) and in the marine biology domain (in the European Fish4Knowledge project) . He obtained a PhD in computer science from VU University Amsterdam in 2001.

Partner 10: Fundación Redes y Desarrollo, DO (FUNREDES)

Network & Development Foundation (FUNREDES – http://funredes.org) is a Non Governmental Organisation with Headquarters in Dominican Republic and legal existence in Paris (France), Caracas (Venezuela), Valencia (Spain) and Cayenne (French Guyana). FUNREDES is a research-action structure, a pioneer ICT for Development NGO (since 1988), member of the APC network, with a large and successful history of project management in the field of ICT4D (REDALC project, a visionary predecessor of the CLARA network; national research networks of Peru, Dominican Republic and Haiti; MISTICA a model project about social impact of ICT; to cite a few). One of the original fields of FUNREDES in ICT4D is linguistic diversity in cyberspace, having conducted early and consistent measurements of the space of languages in the Net and maintaining an observatory of languages and cultures in the Internet.

FUNREDES has a consistent history of efficient collaboration with EC (REDALC/DG13, CARIBCAD/INCO, I-Twinnings/@LIS, WINDS-Caribe/PF7) as well as with UNESCO, ITU, OIF and Union Latina and have a notable record of pushing European vision, especially for collaboration, in Latin American civil society.

Role in DILINET

FUNREDES will offer its experience of languages measurement to assess and evaluate the data resulting of the research. For historical reason with the project, FUNREDES will also maintain an implicit role of cohesion and synergy provider between the WPs and between the partners.

E***xpected outcome from DILINET***

Funredes has been, with MAAYA, the conceptor and promoter of the project, convincing institutional partners to fund the first stage of project definition and consortium creation. Its experience on research-action, indicators production, project management and multi-stakeholder partnership is an asset for a project which original requirement is to merge advanced research with public policy and business players. The success of the project will open avenues for better management of linguistic diversity in the Internet, open the possibility of more efficient search engines, bring new insights in the Digital Divide subject adding to the picture the content divide and contribute to a new paradigm for information society indicators all goals in total coherence with the international mission of FUNREDES as a civil society stakeholder.

Key personnel

Daniel Pimienta, a French citizen born in Morocco, read Applied Mathematics in Nice University and hold a Ph.D. in Computer Sciences. After creating a Software House specialized in APL, he joined IBM France (La Gaude Laboratory) and worked 12 years as Telecommunication System Architect and Planner. In 1988, he joined Union Latina, in Santo Domingo, as Scientific Advisor and Head of the REDALC project for creation of LA&C network. In 1993, he launched FUNREDES and focused on ICT4D, defining and managing more than 30 projects with a vision centred on users and contents and a strong research-action component towards proper methodologies. An active civil society player in Information Society themes (representing civil society voice in the World Summit of Information Society - WSIS) with a special perspective on social impact of ICT, virtual communities and linguistic diversity, he is a member of several ICT4D related global groups such as Francophone virtual university, 3EL, GCNP, EUROLATIS, WINDS-LA, REDISTIC, APC, WSIS-AWARD, UN-GAID and Digital Solidarity Fund. He was given, in 2008, the Namur Award (IFIP WG9.2) for his comprehensive actions in the perspective of a positive social impact of ICT. In coherence with his Funredes assignment, Pimienta has also performed as consultant for UE, ITU, UNESCO, OIF, UNION LATINA, MAAYA and US-AID, among others. Pimienta is a recognized lecturer and writer on the theme related to Information Society with more than 120 conferences (including 25 keynotes) and 70 publications. Pimienta is a member of the evaluation boards of the Journal of the American Society for Information Science and Technology, the Journal of Community Informatics and the Journal of ICT and Human Development.

Alvaro Blanco, a Spanish citizen, graduated in computer science middle degree in La Rioja University. Spend the first few years of work experience around computer related tasks: programming, maintenance, user support and trainer. In 2003 he joins FUNREDES taking part actively in projects SOCINFODO and CARDICIS, and especially the "Observatory of the Linguistic and cultural diversity on the Internet, a languages measurement project started by FUNREDES in 1998. He is responsible for development of the integration of machine translation projects in various conferencing systems, from discussion lists to Elearning platforms such as Moodle. Since 2009, he is the Head of the FUNREDES Branch in Spain.

Both Pimienta and Blanco will also offer services to MAAYA in the frame of this project.

Publications

-Quel espace reste-t-il dans l'Internet, hors la langue anglaise et la culture "made in USA" ?, D. Pimienta in Nord et Sud numériques , Les Cahiers du Numériques, Vol 2 No 3/4 Hermès Numéro spécial sur la fracture numérique, 2001

-"OLISTICA: La búsqueda de maneras alternativas de concebir Indicadores en el contexto de la Sociedad de la Información, D. Pimienta in Proc. del Taller sobre indicadores de sociedad de la información organizado por la RECYT, Lisboa, 6/2001

-The EMEC Methodology (Efficient Management of Multilingual Electronic Conferences) Knowledge Management in a Latin American Virtual Community, D. Pimienta & C. Dhaussy, in Novática, revista de la Asociación de Técnicos en Informática de España; and in Upgrade, digital journal of Council of European Professional Informatics Societies, Vol 3 N1, 2/2002.

- Measuring linguistic diversity on the Internet. A collection of papers by: John Paolillo, Daniel Pimienta, Daniel Prado, et al. . - Edited with an introduction by the UNESCO Institute for Statistics Montreal, Canada . - Montreal: UNESCO, 2005 (CI.2005/WS/06)

- Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, D. Pimienta, D. Prado & A. Blanco, UNESCO, 2009, CI.2009/WS/1

- Digital divide, social divide, paradigmatic divide , D. Pimienta in Human Development and Global Advancements through Information Communication Technologies: New Initiatives, 2011. IGI Global, ISBN13: 9781609604974 - EISBN13: 9781609604981

Partner 11: Vocapia Research, FR (VOCAPIA)

Founded in 2000, Vocapia Research is an R&D company specialized in the development of multilingual technologies for speech and language processing, in particular speech to text transcription systems, audio and speaker segmentation and identification, and language recognition. It has privileged partnerships with the CNRS-LIMSI laboratory.

Statistical methods are used in the VoxSigma software suite to model spoken language and to build leading edge speech processing technologies which can serve a variety of applications, in particular vocal interfaces and automatic audio indexing. Large vocabulary continuous speech recognition is a key technology that can be used to enable content-based information access in audio and video documents since most of the linguistic information is encoded in the audio channel of audiovisual data, which once transcribed can be accessed using text-based tools. Via language identification, speech recognition, and speaker recognition, spoken document retrieval can support random access using specific criteria to relevant portions of audio documents, reducing the time needed to identify recordings in large multimedia databases. Vocapia Research's VoxSigma speech-to-text systems cover many languages.

Vocapia Research and CNRS-LIMSI technologies have been ranked first in the French Technolangue ASR benchmark tests (2005 and 2009) and in the Dutch NBEST ASR benchmarks test (2008). Vocapia Research is providing and further developing these technologies for the Quaero program: the Exalead video search engine has integrated the CNRS-LIMSI/VR text-to-speech technology.

Role in DILINET

Vocapia will offer expertise in speech to text transcription systems, audio and speaker segmentation and identification, and language recognition.

Expected outcome from DILINET

Vocapia recognizes the importance of adapting their technologies to deal with the enormous quantity of heterogeneous data types found on the Web. DILINET will enable Vocapia to have better performing systems and extended language coverage.

Key personnel

Viet-Bac Le is a research scientist at Vocapia Research which he joined in October 2010. He received his PhD degree in Computer Science from the Joseph Fourier University and the CLIPS-IMAG Laboratory, Grenoble in 2006. From 2006 to 2008, he was Postdoctoral Fellow at the LORIA laboratory (Nancy) and at LIG laboratory (Grenoble). He joined the Spoken Language Processing Group at LIMSI-CNRS, Orsay as research associate from 2008 to 2010. Dr. Le has been involved in several speech-related French National projects, the Quaero program and the DARPA GALE program. His research interests include speech recognition, audio segmentation, speaker identification and speech translation. He has over 30 reviewed publications.

Bianca Vieru joined Vocapia Research as research scientist after obtaining her PhD degree in Computer Science from Paris XI University and the LIMSI laboratory in 2008. Her research was focused on the characterization and the identification of foreign accents in French. She received a Master in Cognitive Science in 2004 from Paris XI University and a BSc in Computer Science in 2001 from Bordeaux I University. She has several reviewed publications in conference and journals. She is currently working on the development of speech to text systems in multiple languages.

Cécile N. Woehrling obtained her PhD degree in Computer Science from Paris XI University and the LIMSI laboratory in 2009 after which she joined Vocapia Research as research scientist. Her research was focused on characterizing and identifying regional French accents. She received a Master in Cognitive Science in 2005 from Paris XI University and a BSc in Computer Science in 2004 from Paris XI University. She has published several conference and journals papers on speech. At Vocapia Research she works on developing acoustic, linguistic and pronunciation models for speech to text systems.

Partner 12: United Nations Educational, Scientific and Cultural Organisation, FR (UNESCO)

Since its foundation in 1945, UNESCO as the only United Nations specialized agency for education, science, culture, communication and information, works towards creating the conditions for peace and dialogue among civilizations, cultures and peoples, based upon respect for commonly shared values. UNESCO's unique competencies contribute as well to the realization of internationally agreed development goals.
Under UNESCO's mandate, the Organization also contributes to the building of peace, the eradication of poverty for sustainable development and intercultural dialogue. Through its large network of field offices and National Commission around the world, UNESCO has a comparative advantage to act as a normative setter, catalyst of ideas, clearinghouse and capacity builder within the areas of its global mandate.
Fostering cultural diversity, intercultural dialogue and a culture of peace is one of the strategic overarching objectives that is seen as driving force of development, not only in respect of economic growth, but also as a mean of leading a more fulfilling intellectual, moral and social life. This is captured in the seven conventions in the field of culture and number of recommendations, which provide a solid ground for the promotion of cultural and linguistic diversity. Cultural and linguistic diversity is thus considered as an asset that is indispensable for poverty reduction and the achievement of sustainable

development. At the same time, acceptance and recognition of cultural diversity – in particular through innovative use of media and ICTs, particularly the Internet – are conducive to dialogue among civilizations and cultures, respect and mutual understanding.

Role in DILINET

Within the DILINET project, UNESCO as a standard setter will provide assistance to its Member States to develop a comprehensive approach to language-related policies on the Internet and exploit research results at the national level by providing appropriate tools for the measurement of linguistic diversity on the Internet. The indicators to be developed within the scope of the project will help to formulate language-related policies. The research outcomes of the project will be disseminated and integrated in the international information and knowledge society's institutional frameworks such as WSIS and IGF.

Expected outcome from DILINET

UNESCO considers that cultural diversity and multilingualism on the Internet have a key role to play in fostering pluralistic, equitable, open and inclusive knowledge societies. Within the DILINET project, UNESCO is planning to provide appropriate tools to its Member States needed for the effective measurement of linguistic diversity on the Internet and formulation and implementation of the comprehensive language-related policies and strategies.

Key personnel

Dr Indrajit Banerjee obtained a Bachelor of Arts degree in Arts and Sciences (1985) from the Sri Aurobindo International Centre of Education, Pondicherry (India), and a Master of Arts degree in French (1988) from the Jawaharial Nehru University, New Delhi (India). In 1990, he obtained a Diplôme d'Etudes Approfondies and, in 1994, Ph.Ds in Communication and in Didactics from the Université de la Sorbonne Nouvelle, Paris (France). From 1995 to 1996, Mr Banerjee pursued a Post-Doctoral Research Fellowship in Communication and Media at the Université du Québec, Montreal (Canada). Mr Banerjee was Adjunct Faculty at the University of Ottawa (Canada) from 1996 to 1997. In 1998, he became Associate Professor at the School of Communication and Information, first at the University Science Malaysia and, from 2001 to 2009, at the Nanyang Technological University. Between 2004 and 2009, Mr Banerjee also served as Secretary-General of the Asian Media Information and Communication Centre (AMIC) in Singapore, where he launched a large number of research and publication projects, media conferences, seminars and training workshops. In 2009, Mr Indrajit Banerjee jointed UNESCO as Chief of Information and Communication Technology in the Education, Science and Culture Section and is currently Director of the Knowledge Societies Division, Communication and Information Sector. In this capacity, he has undertaken numerous projects and initiatives, as well as established partnerships with a number of public and private entities. Mr Banerjee is a member of the International Communication Association (ICA), the International Association for Media and Communication Research (IAMCR) and of the Association for Education in Journalism and Mass Communication (AEJMC). He has authored several books and numerous articles in his field of competence.

Dr. Boyan Radoykov joined UNESCO in 1991. He is Chief of the Section for Universal Access and Preservation, Knowledge Societies Division, Communication and Information Sector, UNESCO. He is currently in charge of the Universal Access and Preservation Section and supervises the development and realization of several important aspects of UNESCO's programmes and activities, such as, among others, the Information for All Programme, the Memory of the World Programme, and the implementation of the outcomes of the World Summit on the Information Society. Mr. Radoykov's overall professional career, that has taken him to more than 95 countries worldwide, has been marked by a strong commitment to international cooperation for development. Holder of a postgraduate diploma in economic studies and of a Ph.D. in Political science, he also authored a doctoral dissertation at the University of Paris I, Panthéon-Sorbonne. Member of the French Institute of International Relations, Mr. Boyan Radoykov is author of several books and articles.

Dr Irmgarda Kasinskaite-Buddeberg obtained her PhD in Humanitarian Sciences, Communication and Information specialization from Vilnius University (2006), Master in Information Management (1999) from Vilnius University and Bachelor of Arts degree (1996) from Vilnius Academy of Arts (Lithuania). As a visiting researcher, she conducted research in Lund University (Sweden, 1999), Helsinki University (2000) and Helsinki Technology University (2001). She is currently a programme specialist working at UNESCO's Communication and Information Sector, Knowledge Societies Division in Paris, France. She is in charge of programmes related to the promotion of multilingualism in cyberspace, particularly implementing the Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace. She also works on the projects related to the innovative use of ICTs and Persons with Disabilities as well as development of media and information literacy indicators. She contributed to the implementation of numerous projects, international events, publications and written number of scientific articles on a wide range. Before joining UNESCO in 2002, she worked as Senior Programme Specialist, Department of Information and Informatics, Ministry of Public Administration Reforms and Local Authorities, Lithuania.

Partner 13: Nielsen (NIELSEN)

The Nielsen Company is a leading global information and measurement company that provides clients with a comprehensive understanding of consumers and consumer behaviour. Nielsen deliver's critical media and marketing information, analytics and industry expertise about what consumers watch (consumer interaction with television, online and mobile) and what consumers buy on a global and local basis. Our information, insights and solutions help our clients maintain and strengthen their market positions and identify opportunities for profitable growth. We have a presence in approximately 100 countries, including many emerging markets. We hold market leading positions in many of geographies. Based on the strength of the Nielsen brand, our scale and the breadth and depth of our solutions, we believe The Nielsen Company is the global leader in measuring and analysing consumer behaviour in the segments in which we operate.

We help our clients enhance their interactions with consumers via marketing and make critical business decisions that we believe positively affect our clients' sales. Our data and analytics solutions, which have been developed through substantial investment over many decades, are deeply embedded into our clients' workflow as demonstrated by our long-term client relationships, multi-year contracts and high contract renewal rates. The average length of relationship with our top ten clients, which include The Coca-Cola Company, NBC Universal, Nestle S.A., News Corp., The Procter & Gamble Company and the Unilever Group, is more than 30 years. Typically, before the start of each year, nearly 70% of our annual revenue has been committed under contracts in our combined Watch and Buy segments.

We align our business into two reporting segments, the principal two of which are **What Consumers Watch** (media audience measurement and analytics) and **What Consumers Buy** (consumer purchasing measurement and analytics). Our Watch and Buy segments, which together generated 96% of our revenues in 2009, are built on an extensive foundation of proprietary data assets designed to yield essential insights for our clients to successfully measure, analyse and grow their businesses. The information from our Watch and Buy segments, when brought together, can deliver powerful insights into the effectiveness of advertising by linking media consumption trends with consumer purchasing data to better understand how media exposure drives purchase behaviour. We believe these integrated insights will better enable our clients to enhance the return on investment of their advertising and marketing spending.

Key personnel

René Lamsfuß is Vice President Market Governance & Data Strategy Europe, at Nielsen Europe – leading global provider of information and analytics around what consumers watch and buy. He was until 2011 in charge of Nielsen's syndicated digital product set in Europe before he has taken over his new role. Mr Lamsfuss ensures strong working relationships with industry bodies, joint industry bodies and key clients in order to deliver excellent products that fully comply to privacy and data protection guidelines on both, national and regional level.

Mr Lamsfuss joined Nielsen Online from United Internet Media, the leading German Online Sales House, where he was responsible for market research and media consulting as well as for managing all relations with industry bodies and joint industry committees. In his position at United Internet Media, he was also the Architect of a number of innovative fusion and product development opportunities, such as Targeting.
In 2008 and 2009, Mr Lamsfuss was appointed board member of IAB Europe. He furthermore served as Chairman of AGOF, the German industry body for Online Media Research, where he managed – amongst others - all activities related to privacy and data protection regulation. In September 2011, Mr Lamsfuss was appointed as Chairman of the IAB Europe Online Research Committee, which works together with IAB Europe and local IABs as well as members such as Google, Adobe or ComScore on market governance for online research, especially vis-à-vis decision-makers in Europe. René Lamsfuss holds a MA degree in Geography, Political and Social Science from Heinrich-Heine-University in Duesseldorf.

B 2.3 Consortium as a whole

B 2.3.1 Consortium overview and role of the participants

The selected group of partners participating in DILINET project is highly capable to brightly conduct the tasks associated to this IP, as it has been shown, due to their deep involvement in European and International infrastructures and European Technology Platforms and Technical research capacities.
All of them have been selected because of their own personal and professional reputation on the field of support to R&D computer research expertise, business & linguistic expertise, Internet, ICT Components and systems. Therefore, recommendations from DILINET partners have a high potential to be accepted and adopted to strengthening the technical capacities for supercomputing in Europe guaranteeing the biggest possible impact at worldwide level.
DILINET partners are fully convinced that their participation in this project will significantly contribute to classification of metadata and use supercomputing methods not only for industrial data interests, but on all aspects of the internet including linguistic indicators by identifying strategic partners and by developing international policy objectives and market development priorities, as well as providing a high level competence guidance to assist discussions setting up privilege partnerships.
The DILINET partners are excellently complementing each other covering the necessary perspectives to deal with the most relevant areas of the work programme: Technical experts on metadata analysis and research, business consultancy expertise, ICT industry associations and socioeconomic consultants, as well as industrial & linguistic experts for the applications on pilots.

The consortium represents a well balanced partnership conceived to reach following objectives:

- Collaborative research activities for building indicators for spaces other than the traditional Web and for other approaches than those involving a static vision of existing resources.
- Combination of various research and metadata exploitation methods and analysis to create relevant indicators for complex measurements on the ICT systems
- Application of results on real sectors (business and language) through the pilot projects

All the partners are well-established in their Countries and cover different fields of knowledge. The integration of these peculiarities will ensure the highest level of commitment.
All partners have the necessary competences and in-house resources to carry out the activities planned.

B 2.3.2 Complementarity of participants

Underneath is a table outlining the complementarities of the consortium partners.

Role	Partner	Contribution
Coordinator	GEIE ERCIM	Project management, financial & administrative coordination
Research Partners	CNR	Smart sampling of large-scale data
	CNRS	Language indicators in non-text media
	UPF	Crawling, sampling, network link analysis
	FRAUNHOFER	Sampling, content analysis, system development & integration
	CWI	Data & knowledge representation
Industrial Partners	DIALOGIC	Data collection, user centric measurement, quality assurance
	EXALEAD	Web crawling, web page processing (+scientific coordination)
	VOCAPIA	Speech to text, audio segmentation, language recognition
	NIELSEN	Applications and pilots
Institutional Partners	MAAYA	Societal issues, Dissemination & exploitation (+societal matters coordination)
	FUNREDES	Result assessment and evaluation
	UNESCO	Policies, international dissemination & workshops

Figure 16 - Complementarity of consortium partners

B 2.3.3 Sub-contracting

In order to reduce the initial size of the consortium, it was decided to integrate those partners who deliver a focused subset of activities as sub-contractors of beneficiary 2 MAAYA. The modules, services, and contributions requested from these partners are critical to the project. However, none represent the equivalent of a full task in the work plan, or justify the status of full partner. The sub-contracting option is therefore deemed as being the best alternative to benefit from the expertise and concrete support of these entities.
As a result, MAAYA will establish sub-contracts with four partners: MENON, KYOS, ITU and Union Latine. A description of the specific activities performed by each one of these sub-contractors is detailed in the table below, together with the task it relates to. The total cost of the five MAAYA sub-contracts amounts to 259,000 €, the breakdown of which is provided in section B2.4.3.

Task	WHO	DESCRIPTION
1.4	MENON	MENON will provide a quality assurance plan for all project deliverables, including tools and checklists and a timeline for programmed actions.
3.3	MENON	MENON will produce the first draft of the policy and strategy paper targeted to decision makers as well as the guidelines for use of this document by EU and international organizations.
3.3	ITU	ITU will organize an international conference on the theme of indicators for Information Society and produce a publication of the proceedings.
3.4	MENON	Building on the inputs of the partners involved in T3.4, MENON will produce the first draft of the following documents: o a long-term strategic plan for the DILINET network development, including feasibility, sustainability ad governance elements o a roadmap for future research on the issue of linguistic diversity on the digital

		world o a set of policy and research recommendations beyond the project lifetime.
4.1	KYOS	From the input of the partners involved in T4.1, KYOS will supply the Intellectual Property Rights chapter of the consortium agreement.
4.2	KYOS	From the input of the partners involved in T4.1, KYOS will supply the legal, ethical and regulatory chapters of the consortium agreement.
4.3	KYOS	From the inputs of the partners involved in T4.1 KYOS will prepare the first draft of the document of recommendation for security and privacy matters. KYOS will audit and resolve the privacy and security risks of WP6 using a stress test software.
10.2	UNION LATINE	UNION LATINE with centralize and compile all the terms requiring terminologist definition within the project and translate in the set of languages pre-defined.
11.1	MENON	MENON will contribute to the design of the evaluation framework for applications and participate to the brainstorming for the forecast of future impacts of the project.
12.3	UNION LATINE	UNION LATINE will contribute to the redesign for crawling of the Word Sampling method for measuring languages in the web.

Figure 17 - MAAYA subcontractors

For reference, a brief description of each MAAYA sub-contractor is also provided here:

- ITU

ITU (International Telecommunication Union) is the United Nations specialized agency for information and communication technologies. It allocates global radio spectrum and satellite orbits, develops the technical standards that ensure networks and technologies seamlessly interconnect, and strive to improve access to ICTs to underserved communities worldwide. ITU is committed to connecting all the world's people – wherever they live and whatever their means. Through its work, it protects and supports everyone's fundamental right to communicate. http://itu.int

- Union Latine

Funded en 1954, Union Latine is an International Organization with 35 Member States acting in pro of cultural diversity and multilingualism. One of its programs deals with terminology and languages industry with a focus on the challenges of digitization. http://unilat.org/DTIL

- MENON

MENON is a European research and innovation network, working since 10 years to foster and smooth innovation processes in areas such as education and lifelong learning, international S&T cooperation, knowledge society, social inclusion. It represents a rather unique integration of multidisciplinary expertise and “Innovative thinking”, drawing on a comprehensive and diverse knowledge base combining research methods and their application; network development and implementation; policy analysis and monitoring and valorizing innovative practices. http://www.menon.org/

- KYOS

Based in Switzerland, KYOS is a service-related company sspecialized in the IT security field and in particular in data and network protection, authentication and PKI as well as identity management. KYOS supports companies since 2002 by providing high qualified consultants for projects like vulnerability assessment and testing, system integration as well as network and application security solutions in order to maintain a security level adapted to the client's needs. http://www.kyos.ch/

B 2.3.4 Other countries

FUNREDES in Dominican Republic

B 2.3.5 Additional partners

The DILINET consortium is complete at the time of submitting the proposal. No other partner is foreseen or needed in the foreseeable future. For the national adaptation of the scientific results, several EU and other non EU countries (China) expressed interest in participating in the UNESCO led sub-phase of adapting the language measurement tools developed in the DILINET project (WP3).

B 2.4 Resources to be committed

B 2.4.1 Overview of resources to be committed

DILINET resources have been carefully planned. The project realisation requires mainly high level human resources. The overall planned effort is 775 person-months, or close to 65 person-years, representing a global eligible cost of personnel in excess of 7.5 M€. The project duration is 36 months. This means that DILINET will act as a geographically distributed research team of 21 high level scientists and engineers. The figure below shows the effort allocation over the twelve DILINET work packages, showing that most resources are distributed between the RTD work packages.

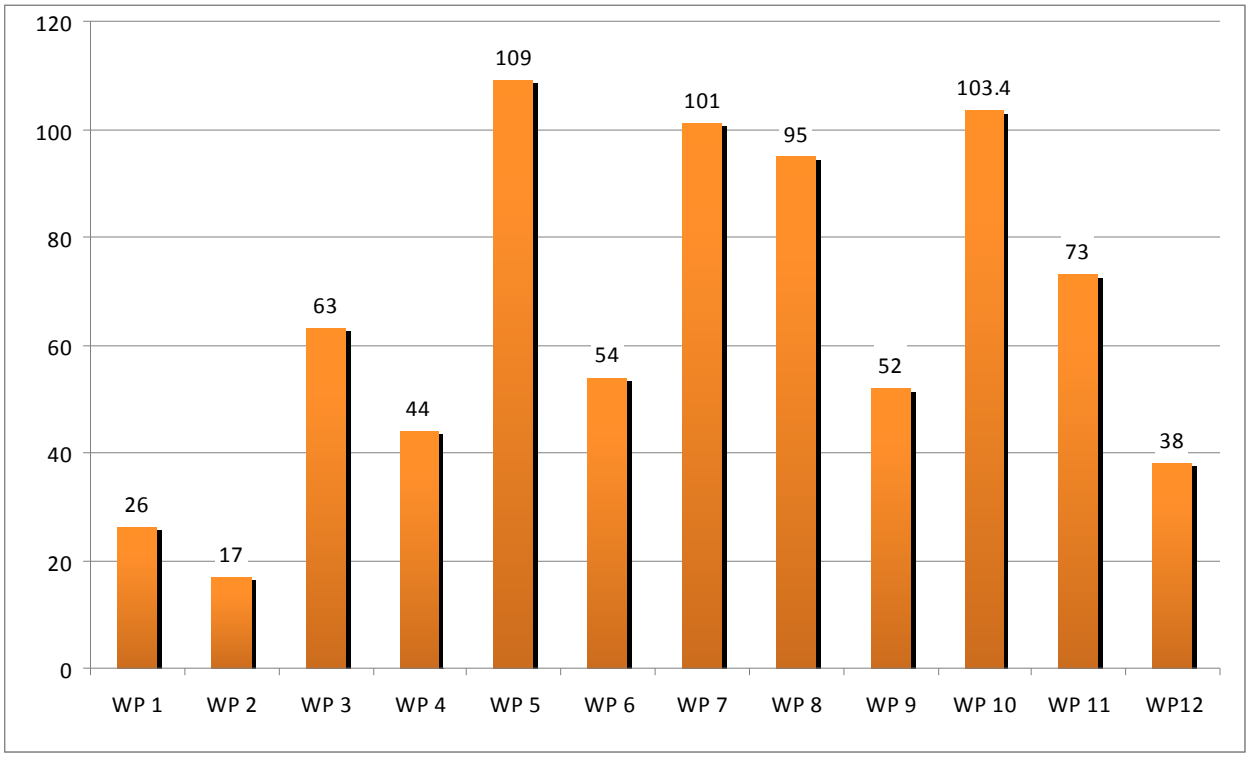


Figure 18 - DILINET effort (PM) allocation per work package

The overall cost structure breakdown into research activities (RTD), management (MGT), and other activities such as dissemination, exploitation and collaboration is presented in the figure below.

Management costs represent less than 4.5% of the total project budget and less than 6% of the total EC financial contribution. The requested EC contribution is **6.520.000 €** representing 74.7% of the estimated eligible costs (budget).

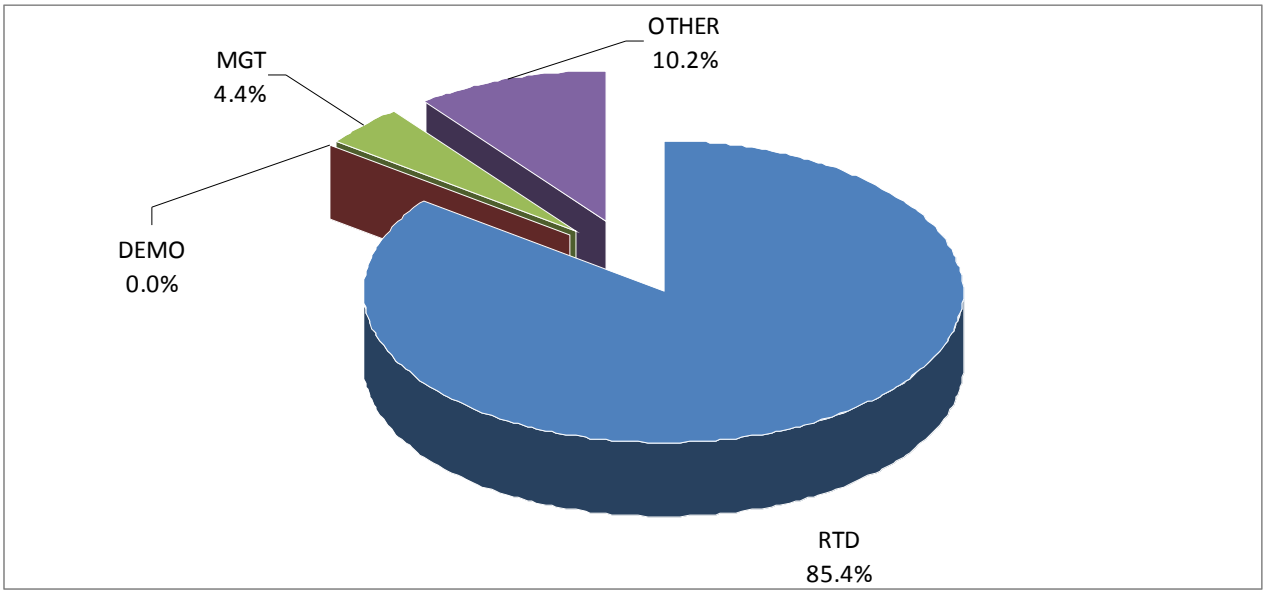


Figure 19 - DILINET cost structure breakdown

The table below represents the consortium budget as it has been entered in the EPSS system. Further explanations are proposed in the following paragraphs.

Proposal Submission Forms

EUROPEAN COMMISSION
7th Framework Programme for Research, Technological Development and Demonstration

A3.2: Budget

Participant number	Organisation Short Name	Organisation country	Estimated budget (whole duration of the project)							Total receipts	Requested EU Contribution
			RTD	Demonstration	Coordination	Support	Management	Other	Total		
1	ERCIM	FR	0	0	0	0	339749	104146	443895	0	443895
2	MAAYA	CH	848981	0	0	0	11425	111381	971787	0	759542
3	CNR	IT	690269	0	0	0	2000	6361	698630	0	526062
4	DIALOGIC	NL	695128	0	0	0	2000	32915	730043	0	556261
5	CNRS	FR	664640	0	0	0	0	41600	706240	0	540080
6	EXALEAD	FR	709200	0	0	0	2000	31200	742400	0	387800
7	UPF	ES	298144	0	0	0	0	33920	332064	0	257528
8	FRAUNHOFER	DE	1213570	0	0	0	4000	10600	1228170	0	924777
9	CWI	NL	593603	0	0	0	2000	8649	604252	0	455851
10	FUNREDES	DO	253080	0	0	0	21060	56160	330300	0	267030
11	VOCAPIA	FR	759776	0	0	0	2000	32160	793936	0	603992
12	UNESCO	FR	51300	0	0	0	2000	390747	444047	0	431222
13	NIELSEN	DE	674460	0	0	0	0	28730	703190	0	365960
		Total	7452151	0	0	0	388234	888569	8728954	0	6520000

Figure 20 - DILINET budget

B 2.4.2 Other direct costs

Travel expenses

In order to contribute to overall spare use of energy and reduction of greenhouse gas emission, travel will be limited and replaced, as much as possible, by e-Meetings (audio/video conferences). A fixed travel expenses budget of 7.200 € is considered for each partner organisation, on the basis of four annual trips to consortium meetings averaging 600 € each (slightly more for FUNREDES who will incur transatlantic travel).

Basic travel costs include participation to 4 consortium meetings (GA and PEB) per year plus additional WP-specific meetings, workshops, conferences especially in the context of dissemination, exploitation, collaboration with other projects.

Other amounts for specific travel have been set aside to cover the following activities:

- Travel reimbursement for the members of the Advisory Board (16 K€), see details below,
- Travel reimbursement for all beneficiaries to attend relevant European and international events and conferences (25K€).

Total travel expenses budget for the whole project is 152.200 €.

Equipment expenses

A budget line of 150 K€ has been established to cover the purchase of the computing and storage facilities that will support the implementation of the DILINET project plan. These hardware resources will be held and maintained at CNR in Pisa.

Together with a small amount of equipment supporting the workshop activity (3.831 €), the total equipment budget for the whole project is 153.831 €.

Advisory board expenses

Advisory board members will not be financially compensated for their contribution to the project. However, in order to ease their involvement, they will be reimbursed their travel expenses.

Assuming 6 members out of 10 or 12 physically participating to the board meetings, 2 meetings organised over the whole duration of the project (end of period 1 and end of period 2) and a unit travel cost of 1000 € for 4 EU advisors and 2000 € for 2 non-EU advisors, this results into a budgeted amount of 16.000 € that has been integrated in the ERCIM budget.

Dissemination material expenses

Workshops will represent a large part of the DILINET dissemination activities. They will be organised by UNESCO with the collaboration of all beneficiaries. To this purpose, an amount of 179.333 € has been reserved on the UNESCO budget for the organisation and delivery of a series of international workshops and trainings (as well as an amount of 70.000 € towards subcontracts to develop and print educational material and to deliver translation services).

ERCIM has also included in its own budget an amount of 7.658 € for various dissemination. This will consist in costs related to the web site(s), project flyers, brochures, printed reports, specific contribution to renowned publications, banners, etc.

Other direct costs synthesis

The total amount of “Other Direct Costs” is 493.022 € for the whole project duration.

B 2.4.3 Sub-contracts

The only sub-contracts foreseen in the project are those under partner MAAYA (P2) and partner UNESCO (P12).

MAAYA

The purpose and scope of the individual MAAYA sub-contracts are described in section B2.3.3; they are quantified in the table below.

Contractors	Specific Activity	Item Cost (€)	Cost per contract
MENON	Plan for quality assurance	15,000	60,000
MENON	First draft of policy and strategy technical guidelines	7,500	
MENON	Design of the evaluation framework, forecast of future impacts	15,000	
MENON	Strategic plan, roadmap and recommendations	22,500	
ITU	International conference and proceedings	30,000	30,000
KYOS	Consulting on IPR matters	18,800	94,000
KYOS	Consulting on Legal, ethical and regulatory matters	18,800	
KYOS	Recommendation/audit for security and privacy matters	56,400	
UNION LATINE	Recognition algorithm and translations	60,000	75,000
UNION LATINE	Redesign for crawling of the Word Sampling method	15,000	
	Total cost of MAAYA sub-contracts:	**259,000**	**259,000**

Figure 21 - Cost detail of MAAYA's subcontracts

UNESCO

UNESCO will subcontract the development and printing of educational material as well as the delivery of translation services for an amount of 70.000 € as part of the project's dissemination activities.

Audit Certificates

Moreover, the average cost for producing Certificates on Financial Statements (CFS) has been evaluated at 2k€ per unit for partners that will require such certificates. These expenses for CFS are declared as subcontracting costs in the management category.

B 3. Impact

B 3.1 Strategic impact

The goal of DILINET is to provide ***an intelligent integrated and comprehensive framework to support decision making*** concerning the use of the Web for horizontal Big data[1] knowledge mining. To this end DILINET will develop new technologies, tools and resources to efficiently crawl, sample, and characterize the web, specifically targeting language use on the web as well as it will build national capacities needed for self-exploitation of the outcomes of the project for strategic planning and implementation purposes.

DILINET responds to a fast growing interest from a variety of stakeholders, including the content industry in charge of producing content for the digital economy, the public sector, and others contributing to the production of online applications and services. The solution to current challenges addressed by the project could eventually go beyond the immediate scope of the project, and lead to a new breed of search engines and help measure non textual content. As such, the DILINET project will contribute to the following impacts from the work programme.

Reinforced ability for a wide range of innovators to tap data infrastructures and to add value beyond the original purpose of the data through data analysis.

DILINET will present an independent and neutral source for the production of statistics on Big Data in web content and help identify which language groups, national sites, and user groups are under-represented in currently available web indices. The project will open alternatives to the current situation, which is characterized by a very limited number of search engines and the lack of a European presence in different areas:

- by adding new possibilities and motivations to existing language dedicated search engines (for example, Baidu in Chinese, Yandex in Russian, Sapo in Portuguese…);
- by opening up research opportunities for innovative ways to crawl which could reach a new breed of search engines capable to recover a wider and un- biaised coverage of the content universe;
- by adding new functionalities to search non-textual contents.

Reinforced ability to find, reuse and exploit data resources (collections, software components) created in one environment in very different, distant and unforeseen contexts.

Google stated[2] in 2008 that its crawler saw a trillion unique URLs, and the number of individual web pages out there is growing by several billion pages per day. But, independent sources estimate that Googles index contains less than 60 billion pages[3]. If social decision making is based upon pages accessible only from Google's (or Bing's or Exalead's) indexes, then more than 90% of the web remains dark. DILINET will identify which language groups, national sites, and user groups are under-represented in these indexes, and provide the means for tapping into these unexploited data sources.

[1] See « two Kinds of Big Data » by Rob Gonzalez, http://semanticweb.com/two-kinds-of-big-dat_b21925, where he distinguishes vertical Big Data (found in large coherent database silos) and horizontal Big Data spread out over the Web. DILINET deals with characterizing the real information found in this second type of Big Data.
[2] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
[3] http://www.worldwidewebsize.com/

DILINET has maximized its potential by bringing together data producers and users:

- It includes researchers and direct users of the research results and ensures a clear perspective of what data are to be extracted from the Web as a Big Data source, and what it will be used for. Indeed, the project was first defined by users who highlighted the need for a better understanding of the nature of the Web. Based on a needs assessment, the project was built around a group of specialized researchers around the theme of intelligent data management for huge data universes, while preserving the particular interests of its users.
- The project includes researchers from various research fields treating the themes of intelligent data management, information retrieval and data mining, spoken language processing, knowledge representation, conceptual maps, complex system management, applied mathematics and statistics. This synergy will provide unique tools for measure actual linguistic data usage and content on the web.

The project will provide the same level of support to minority languages as it will to the most represented languages on the Web.

Much of the information on the Web is not in textual format, and holds data currently unavailable in traditional indices. YouTube, for example, reports[4] that, every minute, 48 hours of video are uploaded onto its site. This corresponds to more than 7 years of video every day. However, none of the audio information in these videos is currently being used to support decision making. To overcome such shortcomings, the speech technologies developed by DILINET will allow decision makers to also exploit spoken data (such as videos).

Value creation through extensive data collection and analysis.

Advanced statistical modelling techniques become accessible to users that have a lower level of mathematical/linguistic training. The content of the web and its temporal trend can be analyzed in a uniform way for a very large number of languages and countries. Especially relevant to decision makers in politics and industry is the consistent elicitation of opinions of web citizens on specific issues across many different languages.

The success of a web-based gram service[5] (which is only limited to non commercial use) is a clear indicator of potential impact of the resources coming out of DILINET. This platform provides access to statistics derived from the language models underlying Bing, and provides a rich data source for many researchers in natural language processing. Since the scale of language models determines in a large part of the success of various statistical natural language processing algorithms, DILINET's application to minority and under-represented languages will increase the number of language tools available for these languages.

Increased economic value of data resources or data analysis through standards for validation, provenance, accountability, access and privacy control.

There are many current projects (TrendMiner, X-LIKE, LiMoSine, LivingKnowledge, etc.) for gathering information from the Web, but DILINET will finally provide the tools for validating the validity of the data extracted.

[4] http://youtube-global.blogspot.com/2011/05/thanks-youtube-community-for-two-big.html

[5] Microsoft provides a public beta n-gram service, to make large scale web language models useable to a wider audience (for non-commercial use only though) http://web-ngram.research.microsoft.com/info/

DILINET's approach to exploiting language use while not revealing any personal data about a user (see WP6) will be a model for privacy protection.

New scientific investigations enabled by large, interconnected data resources and attending infrastructure.

One of the obstacles of developing speech technologies for low e-resourced languages is getting access to sufficient audio and text materials for model training. By identifying what is hidden on the Web, Dilinet will enable researchers and technologists to locate appropriate materials. As current resources on text web data cover only a few languages (e.g. 10 languages for Clueweb09[6]) DILINET will facilitate the combined analysis of web pages in majority languages and a large number of minority languages.

Increased efficiency of organisations and better management of societal challenges (emergencies, planning, ...) through more timely and better decision making.

Analysts will be able to easily compare the public attitude (as reflected in web resources) across languages and regions, enhancing our understanding of uptake of ideas / differences in attitude towards events / policies.

Analysts using the DILINET platform will be able to easily compare public attitude (as reflected in web resources) across languages and regions, enhancing our understanding of uptake of ideas / differences in attitude towards events / policies. The DILINET integrated and comprehensive framework will be used to support decision making process at national level enabling national partners not only to obtain an accurate information about the status of linguistic diversity in the digital world, but also to build sustainable and long term oriented capacities. Many countries around the world do not comprehensive understanding and overview about linguistic resources available on the Internet, therefore adequate language planning and revitalization efforts are limited and not always efficient. The issue was raised at the international level; countries on numerous occasions expressed their wish to assess the language situation in local, national and global content and a need for a tools and resources helpful for the development and implementation of appropriate language policies and measures. Those policies and measures will provide new opportunities for technology to be used for the promotion also of lesser-used languages, establish new standards as well as acquire critical competencies at national level.

DILINET will provide a better view on the actual development of the Web, especially at the fringes (e.g., minority languages). The use of language sheds much needed light on very important global trends such as the loss of cultural identity (e.g., regions in Europe) or the rise of nationalism (e.g., [lack of] integration of immigrants in Europe) or the impact of globalization on culture (e.g., rise of lingua franca).

For instance, some analysts[41] have considered the recent rise of upheaval in Arabic countries was visible in social networks. With better indication of language trends over the Web, such as DILINET will provide, analysis might predict such events rather than just confirming them afterwards.

B 3.1.1 Other impact factors

Market impact

[6] http://lemurproject.org/clueweb09.php/

[41] http://www.slateafrique.com/17731/sur-facebook-arabe-depasse-anglais).

Much required marketing information (e.g., knowing user opinion) is very hard to capture, requiring expensive personal interaction such as calling users, or distributing and collating questionnaires. Thanks to DILINET and its double impact of providing tools for measuring user opinion from horizontal Big Data sources on the web, and proactively from the inclusion of polling structures the voluntarily installed DILINET plug-in, marketing agents will be able to ask specific questions to a wide audience, as well as measure user opinion from real time processing of the Web. DILINET will enhance forecast capacity for content industry in languages hitherto ignored. DILINET will also provide a broad impact on the position of the European market research industry, providing the real time polling tools that will allow EU market research to compete against established US competition: Google Trends, Zeitgeist.

DILINET provides the R&D basis for the strengthening of the European capacities in search engines vis-à-vis their very dominant US competitors. The language dimension emphasized by the DILINET tools will give them a strong advantage on important emerging markets with non-English languages bases (China, Russia, Brazil, Vietnam, India [for Urdu and Hindi]). The extension of web-based Big data mining to audio data will also open new markets.

DILINET methods and results will permit in the future, as an example, to a media industry to decide whether or not invest in producing a digital newspaper edition in a given language establishing more reliable forecast data.

Social impact

DILINET will encompass a systematic effort to guarantee the sustainability of the production of indicators on linguistic diversity and contribute to decision and policy makers' sensitization, as well as to the introduction of linguistic diversity as an important factor for the digital economy.

DILINET will represent a historic breakthrough for the issue of linguistic diversity in the digital world, and contribute to the change of paradigm of the vision of the digital divide. While currently available data are largely limited to understanding the digital divide in terms of *access* to, and *use* of, the Internet, very little is known about the linguistic divide. However, in order to reduce the digital divide, barriers beyond mere access also need to be addressed to give all users the potential to benefit from information and communication. The issue of content and language is fundamental since without relevant content, access itself is useless.

According to the UNESCO Atlas of the World's Languages in Danger about half of the 6,000 languages still spoken today are in danger of disappearing. There is a pressing need in almost every country on this planet for more reliable information about the situation of minority languages, language documentation and new policy initiatives to enhance the vitality of these languages. The new technological solutions and opportunities need to be accompanied with systemic approaches and tools such as to be developed by DILINET integrated and comprehensive framework. It will facilitate formulation of language, education, technology and other relevant policies as well as enable, encourage various stakeholders to take appropriate measures to promote national, regional and lesser-used languages in the digital world as well establish new standards for acquisition of critical competencies at national level.

DILINET will produce new and invaluable information that will not only draw the attention to the opportunities created by linguistic matters but also help identify bottlenecks and shortcomings in today's Web. This will guide policy makers, businesses and users to address the linguistic divide and eventually help to produce more relevant content, and user-centered applications, and to bring more people online.

Through the production of new information society indicators, DILINET represents an important step in addressing the World Summit on the Information Society (WSIS)'s call to "*encourage the development of*

content and put in place technical conditions in order to facilitate the presence and use of all world languages on the Internet".

Increased knowledge on the information society through the production of new indicators

DILINET will open the door to new and user-centered measurement by producing indicators to measure the information society. Since current measurement is largely limited to the quantity of hardware and subscriptions, there is a clear lack of information on the content side, as well as on usage and applications. Eurostat has, with one of the partners of the consortium (DIALOGIC), started to conduct some preliminary research to produce indicators to analyze the digital behaviour or organizations, companies, as well and consumers and citizens. [42] In a similar, but more extensive way, DILINET will challenge the way in which current indicators for the information society are produced. It will work in close cooperation with the *Partnership on Measuring ICT for Development*'s Task Group on Measuring the WSIS Targets, which includes a large number of international parties competent in that field of information society measurement.

Reoriented policies to tackle the Digital Divide

The "*language territoriality*" of the Internet and how it relates to content has often been underestimated in the analysis because people naturally tend to think within their own linguistic boundaries. Yet, it is important to discover and then analyze the *hidden dimension of inclusiveness* of the Internet in order to tackle the coming challenges of its latest stage of evolution, and especially the so much mentioned digital divide. The question about the presence of languages on the Internet may be addressed by crossing currently available Ethnologue data on languages[43] with ITU data on Internet access[44]. It appears that there is a statistical correlation between countries with high linguistic diversity and countries with low Internet penetration[45]. A very important message has been established by the first studies (from FUNREDES/Union Latina as well as from LOP) about the digital divide, showing that the gap of content was one order of magnitude higher than the gap of access for international languages and several orders of magnitude for local languages, as shown in the following figure:

- 4% of global Internet users are from Africa (Source: ITU 2011[46])
- 0.6 % of web pages in French are based in Africa (Source: FUNREDES /Union Latine 2007)
- 0.6% of web pages in English are based in Africa (Source: FUNREDES /Union Latine 2007)
- The percentage of web pages in African local languages vary from 0.06% to 0.0006% depending on the language (Source: LOP 2007)

Much more than currently assumed, the digital divide may be much more an issue of content and language than of access. This is a very powerful argument in favour of digital inclusiveness policies which do not stop at access but, together with access, focus on local content (and indirectly on the education to nurture new content producers, a process which starts by encouraging digital literacy).

[42] See Dialogic, *Go with the dataflow! Analyzing the Internet as a data source (IaD)*, 2008
Available online:
http://www.umic.pt/images/stories/publicacoes1/main_report.pdf
http://www.umic.pt/images/stories/publicacoes1/annexes.pdf
[43] http://www.ethnologue.com/web.asp
[44] http://www.itu.int/ITU-D/ict/statistics/
[45] Other two correlations which are striking matters for thoughts are between high biodiversity and high linguistic diversity and any one of the previous and … poverty. In other word the rich part of the planet is info-rich but linguistically poor
[46] See ITU statistics, at: http://www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html

B 3.1.2 European added value

DILINET's identification of under-indexed and dark language areas in the European community will allow for both the development of region-specific search engines, as well as alerting European language policy officials as to where efforts need to be taken. The European Community has few tools for dealing with the costly language challenge of multilingualism; DILINET will provide inputs for a better management.

In addition, the DILINET aims to reinforce the ideology of multiculturalism and plurilingualism which is seen as a prerequisite for participating in the global information society and in the economic, political and cultural life. International instruments adopted by the Council of Europe such as European Charter for Regional or Minority Languages or recommendations of United Nations demand individuals, groups and various other stakeholders to use languages in every situation of life, including digital world. It should be seen as human right.

B 3.1.2.1 Interaction with other national and international research activities

The project results, especially research outcomes, will be exploited and introduced to the relevant national organizations and institutions. It will enable national partners to use the research outcomes, methodology and training materials for the individual measurement, assessment, planning and other activities directly linked to the language-related public policy formulation and implementation. The indicators will make a direct impact at national and international level, particular providing countries with well-designed and tested tools for linguistic diversity measurement within the WSIS and IGF framework. It will be also used for the policy recommendations, capacity building and research at international level.

B 3.1.2.2 External factors for impact achievement

Research outcomes, methodology and training materials for the individual measurement, assessment, planning and other activities within the project as internal factors will make a directly impact to the language-related public policy formulation and implementation at national level which should be seen as external factor for impact achievement. Efficient national adaptation process is one of the external factors for impact achievement. It is expected that indicators will help to create enabling environment and improve current situation – linguistic and digital divide among countries.

B 3.1.2.3 Standards

To the extent possible the consortium will support EU policies to follow existing standards in international projects and to contribute to new standards in areas of human language technologies.

B 3.2 Plan for the use and dissemination of foreground

B 3.2.1 Dissemination

The DILINET project has started from the requirements of indicators from a group of users composed of International and civil society organisations which have sought the support of a distinguished set of multi-disciplinary researchers capable of overcoming the actual limits of the state of the art to understand the structure of the web contents and its meaning in terms of various indicators of the

evolution of the Information Society (especially in terms of language, a key element to gauge the content divide, the very subtract of the digital divide).

It is then logical that the exploitation of the results of the research occupies an important part of the project and the dissemination of the results answers the characteristics of a wide multi-disciplinary range of research around the widest and fastest growing data set as of today, the World Wide Web.

For Scientific Partners

The project will naturally contribute to the advancement of science by publishing. Project members are expected to produce publications at major scientific journals, and present their work at international conferences and workshops. The project will also organize special sessions at major conferences, and satellite workshops collocated at major conferences, for maximum visibility.
Beyond the common scientific process, the consortium will participate in international benchmarking exercises and campaigns to compare the advances and innovations of DILINET to the state of the art.

For the consortium

The DILINET project considers it essential to communicate its results to the open public as well. This activity will involve communicating the goals and activities of the project in lay-men's terms and to reach out to the public and to public bodies. Some possible mechanisms are Public demos and Press Conferences and Participation in Trade-Shows and Fairs.

Furthermore, the marked interest of International Organisations such as UNESCO, ITU, OIF or Union Latine, as well as the linkage with international cooperation frameworks such as WSIS and IGF, will deliberately orient the exploitation of the results of the project towards public policies in the field of Information Society, be in terms of Linguistic Diversity, be in terms of Digital Divide; all that without reducing the possible exploitation of the project results towards a contribution to the most sensitive application of the Internet, the Search Engines, in terms of economy, culture and even information ecology; and non withstanding the effective contributions of the project in terms of digital economy, more precisely the content industry and its almost unsolved requirements as of today in terms of forecasting marketing niches linked to languages.

Beyond the traditional dissemination channels in the form of scientific publications in the various fields where DILINET will bring significant contributions, this work package will offer a structured set of activities linked to UNESCO mandate in the area of international cooperation. The concept is to use the research results, methodology developed and experiences gained during the project implementation period and make it more accessible, adaptable and replicable for other organisations such as national governmental organisations responsible for statistics, relevant ministries responsible for information policies and language issues, higher educational organisations and other institutions. Within the framework of DILINET, the emphasis will obviously not be put on the international cooperation itself, but rather on the distribution and application of the scientific results targeting national institutions and building their capacities to use developed indicators for own planning purposes.

Hereafter is a proposed broad list of activities to be implemented in close cooperation between the partners with a leading role of MAAYA, a major contribution of UNESCO and ITU and the participation of other international organisations and DILINET partners:

Development of training materials

- The products and methods developed and assessed by DILINET will be used for the development of the training materials for higher educational institutions in Open Educational Resources format.

- UNESCO, together with other partners, will develop training resources on measurement of Linguistic Diversity on the digital world.

Capacity building
of national governmental organisations on application of the Linguistic Diversity in the digital world.
The developed and tested indicators will be used for the capacity building of national governmental organisations responsible both for the national statistics and information policies. This includes:

- Consultations on how the methodology and tools developed could be used at the national and international levels by other stakeholders (the scope will depend on which outcomes/documents will be made public).
- Development of Technical guidelines on how indicators and methodology could be applied by other bodies such as national statistics departments and governmental organisations responsible for information policies.

International events
Several events at the international conferences will be organised to present the outcomes of the project and share experience with other experts working in those fields, few under the responsibility of UNESCO on the subject of linguistic diversity (WSIS Forum(s) and IGF); one under the responsibility of ITU, on information society indicators. Whenever schedules allow, those events will be linked and/or integrated to the celebration of the WSIS+10 events in 2014.

Integration of the indicators
Integration of the produced indicators into the other existing Information Society indicators schemes in connection to the WSIS action plan, specifically the Workgroup on Measuring WSIS Target composed leaded by ITU and composed by EUROSTAT, OECD, UNESCO/UIS, UNCTAD, UNDESA, UNECA, UNECLAX, UNESCAP and UNESCWA.

Sustainability of the indicators
This work package will take the required actions to avoid that the results constitute only a one time shot and then organize the necessary steps in order to obtain a sustainable production of indicators.

Research roadmap
Beyond the described dissemination and exploitation of the results of this major work package, which will represent a keystone and immediate return on the research investment, a roadmap of research will be designed for the main subjects organised around intelligent data management. This research roadmap will pave the ground for a fruitful cooperation between European Union scientific and industrial entities and international and civil society organisations, in the framework of a set of scientific matters representing the edge of the digital challenges.

B 3.2.2 Exploitation Plans

Beyond the exploitation plans outlined in WP3 which are mainly oriented towards public policies for language on the Internet and information Society indicators, and beyond the set of applications designed in WP11 to demonstrate concrete exploitations of the researches results and products, DILINET opens a wide range of exploitation opportunities in the public and private sector. The main aspects are shown in table 22 and will be described in depth in this section.

Market Size	Text analytics:	835 million USD
	Market Research Europe:	13.3 billion USD
	Speech Recognition:	13.6 billion USD

Market Requirements	Analysis sources:	blogs 62%, news 41%, reviews 30%
	Analyses methods:	topics 83%, opinions 77%, named entities 76%
	Analysis questions:	Brand manag. 39%, voice of customer 39%
Public Target Organizations	Political Organizations:	Opinions of citizens across languages
	Language Monitoring:	Monitor language use
	Public Health:	Patient feedback evaluation across languages
Private Target Organizations	Media Companies:	Multimodal customer feedback evaluation
	Market Research Firms:	Multilingual product/reputation research
	Advertizing companies:	Multilingual marketing customization
	Search Engine providers:	Joint multilingual indexes for text and speech
Dilinet Selling Points	Web sampling:	Representative text / audio web sample
	Personal content:	Integrated panel of volunteers
	Multilingual:	Consistent analysis for very many languages
	Text and speech:	Integrated analysis for text and speech

Figure 22 - Main aspects of DILINET exploitation

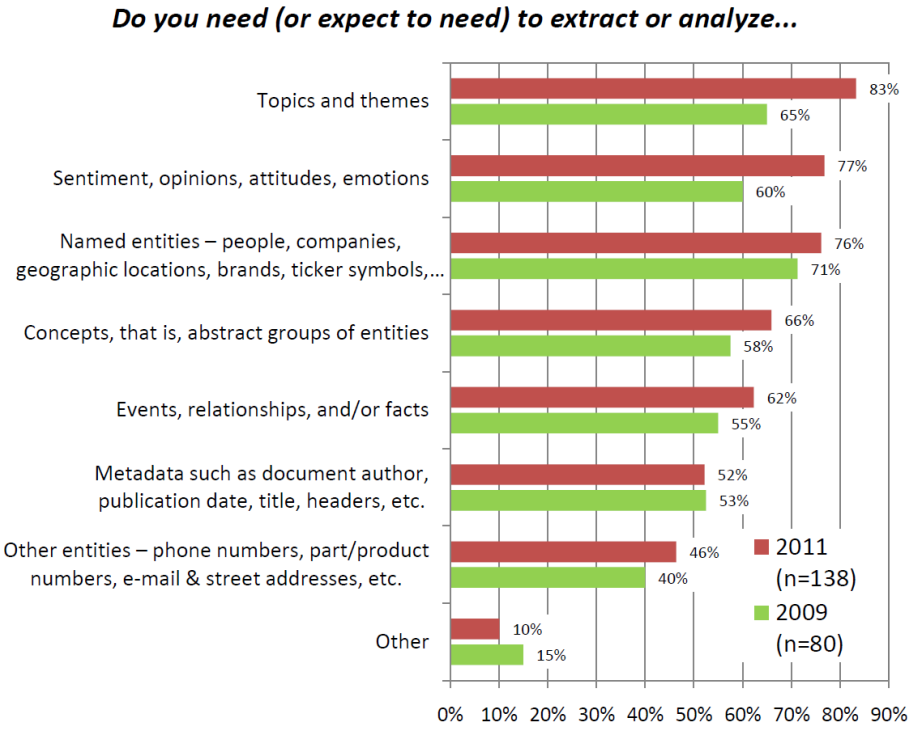
B 3.2.2.1 Market Size and Characteristics

The addressable market for text/content analytics is large. According to Information Week[47] the revenues for text analytics (Software and service) were 835 million USD in 2010. The report predicts an annual growth of 25%-40% in the coming years. Text analytics in part belongs to the field of market research which in 2009 had sales of 13.3 billion USD in Europe and 30.4 billion USD globally[48]. Another market addressed in DILINET is the market for automatic speech recognition which (together with text-to-speech software) has a size of 13.6 billion USD in 2010 and an estimated size of 18.9 billion USD in 2015[49].

A number of factors contribute to sustained growth of text analytics, foremost the growth of social platforms, which have become essential life tools for individuals as well as an important personal and political communication, research, business marketing and commerce channel. Keeping up with social networks and internet content is a must for every consumer-facing organization, and automated monitoring, measurement, and engagement is the only way to deal with Social Network's variety, volume, and speed. DILINET's does this uniformly for a large number of countries, languages and cultures allowing to treat Europe as a single community and market instead as a patchwork of single countries. Content analytics makes sense of rich media. The technology finds and exploits patterns – what's in a given piece of content and how the content of content changes over time – with DILINET simultaneously addressing text, speech and video. Market Requirements on Content Analysis
In 2011 Altaplana conducted a survey of the text/content analytics market[50] and compared it with their 2009 study. Most important information analysis targets for the respondents were social and online

[47] InformationWeek, May 12, 2011, http://www.informationweek.com/news/software/bi/229500096?pgno=1
[48] ESOMAR Industry Report (2010): http://www.esomar.org/uploads/industry/reports/global-market-research-2010/ESOMAR_GMR2010_Cover-Contents-FirstChapter.pdf
[49] BCC research, http://www.bccresearch.com/report/voice-recognition-tech-markets-ift039b.html
[50] Seth Grimes: Text/Content Analytics 2011: User Perspectives on Solutions and Providers. http://altaplana.com/TA2011

sources (blogs and other social media: 62%, news articles: 41%, on-line forums: 35%, review sites: 30%) as well as direct customer feedback (customer/market surveys: 35%, e-mail and correspondence: 29%). DILINET collects a representative sample of the internet and allows investigate all sources according to their relative weight including audio and video sources. In addition DILINET covers parts of the Internet not covered by the search engines as well as the Deep Web and therefore offers a more comprehensive and representative view then prior analyses.



Source: Text/Content Analytics 2011: User Perspectives on Solutions and Providers. p.31. http://altaplana.com/TA2011

Figure 23 - Altaplana 2011 survey of the text/content analytics market

As shown in figure 23 the most important analyses the respondents were looking for are "Topics and themes" (83%) and "Sentiment, opinions, attitudes, emotions" (77%), which are covered by DILINET for multiple languages.

The respondents also detailed the properties of a good solution: "Broad information extraction capabilities" (63%), "Ability to use specialized dictionaries, taxonomies, ontologies, or extraction rules" (57%), and "Deep sentiment/emotion/opinion extraction" (57%), "Support for multiple languages" (44%) as well as "Big Data Capabilities, e.g. via Hadoop/MapReduce" (29%). DILINET advances uniform multilingual analyses emphasizing content category classification and aspect-aware opinion mining. The top business applications of text/content analytics identified by the respondents were: "Brand / product / reputation management" (39% of respondents), "Voice of the Customer / Customer Experience Management" (39%), "Search, Information Access, or questions Answering" (36%), "Competitive intelligence" (33%).

Finally the respondents expressed a strong tendency to analyze languages different from English within the next two years, e.g. French (62%), Spanish (54%), German (46%), Chinese (46%), but also more exotic languages like South Asian languages (Hindi, Urdu, etc. 19%) or Bahasa Indonesia/Malay (5%).

B 3.2.2.2 Target Organizations of the DILINET Exploitation

The target organisations for the DILINET exploitation can be grouped into two main types, the public and the private sectors. The following descriptions provide an illustration of several of the relevant organisations in public and private sectors which will be targeted:

Target Organizations in the Public Sector

- **Political Administrations and Organizations**: A number of recent political situations show the value and effectiveness of Web 2.0, in relation between citizen and public administrations. This is the case of different electoral processes (i. e. the Presidential Election in the United States of America, the current primaries in the US, as well as some elections in European countries); also in legislative processes (the increasing use of tweets and Facebook encouraging to lobby for or against a specific legislation being under discussion), or in spontaneous citizen movements such as it happened in the explosion of Arab spring or in the various occurrences of 'indignados'. The web is a tool for free thought and for free speech, and Web 2.0 has accelerated and democratized the ownership of information by the citizens. Being able to tap into this massive knowledge and opinion base would allow administrations and political parties to better address the needs and wants of their constituents with fewer resources. For European governments and institutions it is especially valuable to elicit the opinion of European citizens in a consistent way. The push in administrations for Open Data will also open windows of creative opportunities for the exploitation of the DILINET platform.
- **International Language Monitoring Organizations**: The indicators developed in DILINET may be used continuously to monitor policies set by international normative instruments, various international forums and programs (***European Charter*** for Regional or Minority ***Languages*****,** UNESCO ***Recommendation concerning the Promotion and Use of Multilingualism*** and ***Universal Access to Cyberspace,*** Millennium Development Goals, WSIS, IGF, Broadband Commission, etc.). There are a number of additional societal activities which may be funded by international organizations:
  - Language Revitalization: Customized design of revitalization programs for language based on their virtual presence or in function of thematic gaps or language deficiencies in the Web (i. e. lacks of terminology or phraseology, spelling differences, incorrect coding, etc.). These programs include the implementation of training seminars, publications, teaching or advocacy, demonstration videos, etc.
  - Design of programs on digital literacy following of specific deficiencies of each language or region.
  - Design of adjustment programs for media use by the people according to cultural preferences and constraints: i. e. supports voice and/or video for oral languages
  - Corpus management, from mostly rare languages, to create educational materials, prevention materials (diseases, disasters, etc.), awareness materials (rights of women and children, regional and minority language users, etc..), information materials (administrative, legal , etc.).
  - Design of training programs for journalists using different media (blogs, Internet radio and TV, etc.) for regional or minority languages.
  - Creation of repository(s) of open source software in different languages and policies of translation into lesser-used languages.

- **Prevention and Health:** As widely admitted, the prevention of disasters, disease, and crime are topics which are globally relevant. Many health care providers, such as national health systems, hospitals and assisted care facilities are searching for ways to reduce spiraling costs of care. National systems of security, insurance companies, justice and other state bodies or private entities for prevention need rapid interaction with the citizen. An advanced feedback evaluation based on the opinions of citizens and professionals and performed with the techniques of the DILINET project would assist in reducing costs and improving standards of care and prevention.

Target Organizations in the Private Sector

- **Publishing/Media Companies**: Worldwide, the publishing industry is one of the industries which are suffering the worst under the current economic crisis. Some publishing houses have integrated Web 2.0 features such as blogs, forums and twitter into their existing legacy systems. This allows for customer/reader feedback. This creates a problem of measuring and analyzing the feedback, which is not the core business for the publishing companies. The platform developed in the scope of the DILINET Project would give publishing companies the basis technology to easily measure and analyze feedback from their customers on the basis of blog comments, forum entries and/or tweets (Twitter) in a uniform way in different languages. This would then give publishers the ability to tailor their content offering to what their customers/readers find interesting, identify experts from the community and generally engage their customer base in new and interesting ways.
- **Market Research Firms**: Traditionally, market research firms have relied on a variety of techniques to record, aggregate and summarize qualitative and quantitative data. The problem for international markets is the fragmentation of the research by language barriers. Within the scope of the DILINET project, multilingual techniques will be developed and applied to properly weighted random samples of the web. The technology will yield comparable results for the different languages (both text and audio) thus providing a unified view on the perception of a product or brand. Another problem which the project can solve for market research companies is that of data volume. Such companies can already sift through the mountain of data available on the web, but at a large cost due to the fact that it is mostly entered manually. Thus, the DILINET project allows for quicker and more accurate collection, preparation and analysis of information for market researchers. As Nielsen, the globally largest market research enterprise is a partner of the project, the DILINET approach will especially be geared to up to date and relevant solutions.
- **Advertising and communication companies**: The advertising and communication sector has been hit especially hard during the past economic downturn. The DILINET project will enable companies and their customers to benefit from multilingual analysis modules, allowing advertising and communication companies to see what is being said online about their customers and then customize ad placement, marketing campaigns, brand development and slogans. Furthermore, companies can be alerted to problems such as online brand bashing, where unfounded rumors about their products or services are posted online. Yet another facet is product development, where communication companies can work with their customers to help develop products or features of existing products based on input obtained from the DILINET platform.

- **Search Engine Providers:** DILINET will produce a new measure of the content and makeup of the Internet. We have mentioned that one, large untapped source of information on the Internet concerns the currently ignored content of audio-visual streams. Current search engines index videos only by the meta-data associated with audio data. The advances in audio processing, across many languages, that will occur in DILINET will allow search engine companies to create a new generation of search engines including the language factor in both textual, as is done today, as well as in audiovisual searches, currently ignored. The new geography of the Internet that will be brought to light will also illustrate where search engines can profitably devote energies to index unexplored content brought to light by DILINET.

Another target group is the IT industry itself, which can use the functionality of the DILINET platform and data processing modules for application development.

B 3.2.2.3 DILINET Selling Points

The DILINET solution has the following **unique advantages**:

- The results are based on a representative sample of web pages or a representative panel of web users. They do not depend on the unknown crawling and indexing strategies search engine providers. Very important is the coverage of the Deep Web and web pages ignored by usual search engines.
- DILINET provides an integrated coverage of complementary personal restricted web content (email, social networks) by an integrated panel of individual volunteers.
- The analysis of contents and opinions is done for very many languages in a uniform way yielding comparable results covering many countries.
- In the same sampling scheme text, audio and video sources are covered and analyzed. Therefore the results for each modality can be combined in a consistent way and provide a unified picture of web content. For instance, we are not aware of opinion mining and content analysis for many languages integrating the results from text and speech analysis.
- The DILINET platform will provide a cost-effective framework for the highly parallel, modular, and fault-tolerant execution of sampling and analysis modules collecting and processing many hundred Terabytes of web pages, audio and video files.

Finally a very important aspect is that scientific and commercial partners of the DILINET project, especially Nielsen - the globally largest market research firm - have an outstanding reputation.

The DILINET consortium can financially profit through different **revenue streams**, such as:

- The conduction of custom analyses for business partners, e.g. by estimating the opinion of all European citizens on political issues (e.g. "genetically modified food") or on product brands. This also covers continuous language monitoring for international organizations.
- Licensing of additional modules, which extends the functionality and capability of the platform. These modules will be proprietary, and will not be under an open source license. For example, better monitoring and profiling modules for the system;
- The development of additional applications based upon the core functionality of the DILINET platform which would benefit and extend the consortium members' own line of business;

- Data as a Service (DaaS): Cloud application developers can use existing data which is provided by other resellers. Data can be licensed as a basis for applications. For example, a company can harvest forum content via web-sampling, process and annotate it with additional meta-data such as content categories, detected named entities, geo-annotations etc. This annotated forum feed can then be subscribed to and licensed by other companies which, for example, develop applications for "political trend detection" based on web community observation.
- Data sources, such as multilingual new feeds and blogs which are enriched with meta-data;
- Consultancy services, which can be offered in conjunction with technical support and development.
- Application Service Provider: Hosting of the platform and second level support.

Finally, as this platform will be developed with open standards in mind, this will serve as a basis for other companies to develop additional technological and sector specific modules, creating a rich ecosystem of business solutions increasing the overall value of the platform. This will increase the importance and reach of the DILINET platform.

Much of the infrastructure of the DILINET platform will be based on existing open source components. The architecture will need parallelization frameworks like Hadoop[51], Mahout[52], or Storm[53] which are under the Apache 2 license or the Eclipse Public License. This will constrain the open source license of the part of the DILINET platform, which will also be published as open source.
As this area is constantly changing, open source licensing experts will be involved in the project, such as lawyers specializing in open source, and members of the open source community. Related issues which must also be investigated include:

- Developing a Intellectual Property Rights (IPR) strategy (see also Section 3.2.3);
- Developing a Licensing strategy which fits within the IPR strategy;
- Licensing agreement for software components to end-users or value-added distributors.

B 3.2.2.4 Particular exploitation targets and plans of the project partners

The partner consortium has strong ties with contacts in large, international media conglomerations, leading business enterprises as well as in the public sector in areas such as healthcare, government and academia. The DILINET platform will provide services to a wide range of public and private entities, as well as consumers. Project partners include market leading organisations in their respective fields.

3.2.2.4.1 Fraunhofer

FHG (Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.) is Germany's leading organization of research institutions for applied research. They are also currently involved in numerous industrial projects in the fields of machine learning and data mining. FHG's customer base includes large corporate entities such as Deutsche Post AG, as well as municipalities such as the city of Cologne. FHG is carrying out contractual research work for many major companies in Germany. As an applied research organization, the core mission of FHG is to offer the DILINET technology to these companies, thus promoting the uptake of intelligent information management technology in commercial settings. FHG is planning to exploit the content mining technology developed in this project in its commercial projects, in particular in fields that face the problems of large and distributed amounts of data. In particular, FHG

[51] hadoop.apache.org
[52] mahout.apache.org
[53] https://github.com/nathanmarz/storm

will investigate the use of DILINET's results in existing collaborations in the business areas of online advertisement, Smart Semantics, and fraud detection.

3.2.2.4.2 Exalead

Exalead recognizes that the proportion of web-based information that is only found in audio streams is growing radically. As a company providing search engine technology of search based applications, Exalead has already begun to build a search solution for audio-based information. Its demonstrator Voxalead[54] currently indexes audio streams for nine languages, using DILINET partner Vocapia's technology. Exalead is planning to exploit the extension of languages that Vocapia will produce during DILINET to further enrich its Voxalead offering. Exalead will also examine whether the country and region specific search engines built in its application section are commercially viable, or will be financially supported by the targeted countries or regions.

3.2.2.4.3 Vocapia

Vocapia Research develops leading edge speech processing technologies. Dilinet will allow Vocapia to extend their offering of languages, so as to cover a larger potential client base. Vocapia expects to improve their language and dialect identification technology in particular with the identification of languages in multilingual documents.

3.2.2.4.4 Nielsen

The consumer industry needs to reflect the changing communication behavior of customers in their complete business models. Then industry needs to include this in product and service innovations processes, marketing, information and advertising and the communication process with the consumer. The consumers are the new authorities in the digital era. User generated media is becoming the content of the internet and it is important for the industry to understand and listen to the consumer.

DILINET enables the industry to understand the consumer so that the industry is enabled to identify trends, demands and wishes of consumers and to provide the related products and services. The digital media is especially supporting small and medium enterprises to be innovative and creative and DILINET will provide them with relevant information for their business.

Nielsen plans to exploit the DILINET system and its analysis tools for

- Trend analysis
  - Analysing new ideas and trends in the digital community to identify new market places and requirements
- Product and service development
  - Develop new product and service ideas directly based on consumer requirements
  - Identification of purchase drivers
  - Identification of market sizes
- Improvement of product and service quality
  - Analysing key factors for product and service quality
  - Tracking system based on consumer feedback
- Product and service tracking
  - Getting insights into sentiment of product and services
  - Tracking of issues and challenges

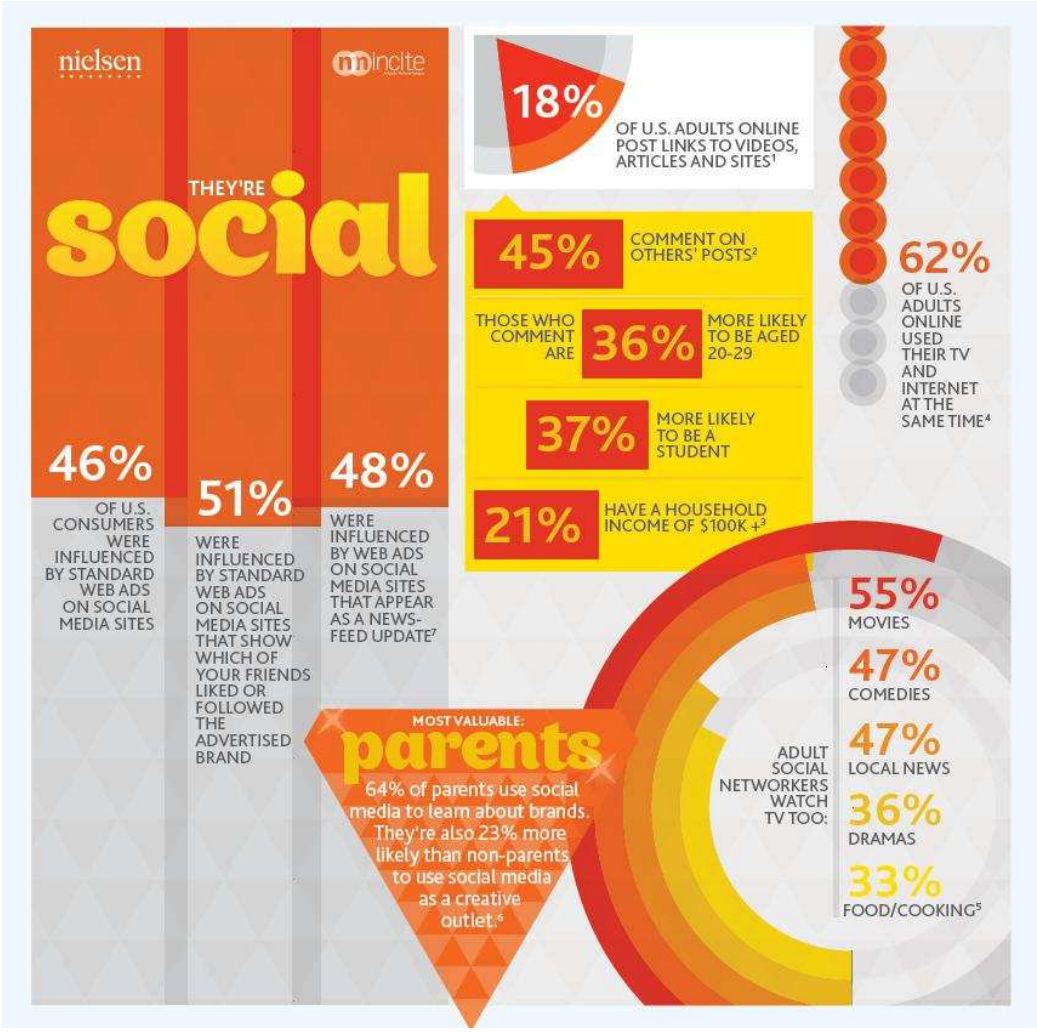
The markets have changed from seller's to buyer's markets and the digital technologies are forcing this. For instance the Smartphone revolution is bringing consumers in the position that they are able to check product prices within the store based on the barcode as well as they could send information about a

[54] http://voxaleadnews.labs.exalead.com

good product offer or purchase advisory service directly to their friends via social networks. New technologies are forcing new communication processes that the industry needs to convert in market intelligence, product and service innovation, marketing and communication strategies.

Social, local and mobile media seems to be driving much of the conversation about online opportunities. But at the end of the day, there is only one constant common denominator across the Web: the consumer. The understanding of the consumer is absolutely necessary for the industry. This was the case in the past and it will be the case in the future. Nowadays the industry needs to include the influence by social, mobile and local experiences online as a vital part of this understanding.

The below picture 24 from Nielsen and NMIncite highlights digital consumer behaviors and consumption patterns that can help brand advertisers understand their most valuable customers and how they're engaging across social, local and mobile.



SOURCE INFORMATION:
1. Nielsen, @Plan (Release 3 2011). Social Networking Activities, Online 18+
2. Nielsen, @Plan (Release 3 2011). Social Networking Activities, Online 18+
3. Nielsen, @Plan (Release 3 2011). Social Networking Activities – Comment on others postings, Profiled by Demo
4. Nielsen, @Plan (Release 3 2011). TV/Internet Concurrent Usage (Yesterday) – Used TV/Internet Concurrently (Yesterday), Online 18+
5. Nielsen, @Plan (Release 3 2011). Social Networking Activities – Any Social Networking Activity, Profiled by TV Programming Viewership
6. NM Incite, State of Social Media Survey (April 2011). Parents = Have Children <18
7. Nielsen, Global Online Survey (Q1 2011)

Figure 24 - Nielsen Findings on Social Networking Activities in the Internet

This combination of academic and industrial partners is perfectly suited for working with existing customers and collaborators in a variety of industry and academic segments to develop the DILINET platform for market use. Suitable targets will be defined in the early stages of the project. As each target has specific needs which can be met through the technology developed in the DILINET project, partners in the project consortium will be involved in cultivating and extending ties to their existing customer base to keep these key assets informed of current and future project developments. The project partners will also organize dedicated demonstrations of running prototypes and scenarios of use, i.e. extensions of the use cases, for key persons in the defined target organisations.

B 3.2.2.5 Overall Marketing Strategy and Perspectives

To ensure the sustainability of the DILINET project, consortium partners will formulate a detailed overall exploitation plan. As presented above, the industry partners in the DILINET project already operate successfully in their respective markets and sectors. This deep industry understanding will be drawn upon to assist in marketing the results and achievements of the DILINET project. The exploration and dissemination of the DILINET project will be based on modern marketing and communication best practices. Marketing and communication activities will be divided into the following:

- Online
    - At the heart of every modern marketing and communications plan is the design and implementation of a web portal. The web portal severs to inform target groups of the purpose, achievements and goals of the research project. It will be structured in such a way that the information needs of the target audience are met;
    - Special attention will be given to a media section on the web portal where presentations, press releases, case studies, success stories, etc. will be offered as downloads. Users will also be given the opportunity to share and recommend these documents via social bookmark services such as delicious, Mister Wong, etc;
    - Interested parties will also be given the possibility of registering for updates, such as press releases, newsletter and announcements about conferences or speaking engagements pertaining to the project;
    - Web 2.0 technologies will also be incorporated into the plan. Services, such as professional social networks, document sharing platforms (i.e. SlideShare.net), blogs, etc. will be used to extend the reach and market presence of the DILINET project and its results;
    - Publishing of technically oriented whitepapers which target technical managers, software architects and developers. These papers can easily be understood and describe the principles and potential of the DILINET platform.
    - Identification and visits of appropriate national and international trade fairs.
    - Early development of a web-based Demonstrator(s), which will serve to showcase the developments and capabilities of the DILINET platform.
- Offline
    - As stated above, the definition and exploration of existing business contacts of the partners. Special attention will be given to multipliers (i.e. persons or businesses which will be champions of the DILINET project);

- Definition of key industry exhibitions, conferences, and technology fairs, such as software developer and open source conferences; Presentations of the Demonstrator(s) at exhibitions, conferences, and technology fares;
- Constant engagement with the press.
- Definition of appropriate technical and industry journals in order to engage with editors toward the goal of publications of articles, features, and editorials describing the system, the results of the research and the benefit for public and private entities
- Development of standard print marketing materials, such as one-pagers, project flyers, and a standard presentation (i.e. Power Point) which can easily be adapted to serve the needs of the consortium partners.
- Development of success stories and case studies.
- Grass roots activities, such as participating in local and regional Meet-Ups or BarCamps, speaking engagements at universities, etc.

Further marketing and business development activities will be geared toward the identification of strategic partners in the subsequent geographic targets.

There are several options for market entry strategies for the DILINET project. These include, but are not limited to:

- The definition of appropriate targets, both public and private entities, and partners in the network of the consortium partners who have an interest in the positive outcome of the DILINET project, and are also suitable for obtaining first experiences and willing to be used as success stories;
- Building out additional use-cases, if need be, to fit the needs of the defined targets;
- Definition of national and/or regional distribution partners or value-added resellers and independent software vendors (ISV) with a minimum direct involvement;
- Development of licensing agreements for the national and/or regional distribution partners or value-added resellers;
- Strategic partnerships with established players in the market.

These strategies are not mutually exclusive, but can be combined and mixed with - based on geographical difference - and evolve, depending on market response and first experiences. Additionally, as each project partner has key insights into their geographic region and industry sectors they serve, the market entry strategies will be adapted accordingly.

B 3.2.3 Management of intellectual property

Intellectual property comprises all kinds of intellectual property and know-how connected to the project that will be generated during the project and as a result of the same. Intellectual property further comprises participants' pre-existing intellectual property and know-how owned by the participants before the start of the project, and also intellectual property and know-how created outside of the project, during its duration, and which is connected with the project.

Proper intellectual property protection will be considered, also under the perspective of possible copyright protection, patentability and any other king of intellectual property protection, in relation to software and more generally any kind of know-how that will be produced in the course of the project, as

a result of the same or that will be comprised in the outputs of the project. To this regard, know-how and any kind of intellectual property right developed in relation to a specific stage of the project or as an output of the same will be protected through appropriate procedures and agreements established among the members of the Consortium.

Management and protection of knowledge and intellectual property will be eased, within the project, by the tight interaction with the legal partner. From the early stage until the end of the project time, the legal department of the partners will provide continuous assistance as to management of the knowledge produced and protection of intellectual property rights in any way arising or connected with the project. The partners will determine the appropriate knowledge management procedures and rules within the Consortium at the various stages of work and thereafter, especially for what concerns the innovation aspects of this project.

The Consortium Agreement will devote specific and significant attention to the issue of intellectual property rights management. In the Consortium Agreement they will be defined and specified procedures and rules for a proper handling, ownership, managing, protection and granting of the knowledge and of any relevant intellectual property rights, in any way produced and of any kind, with regards to both internal usage for scopes within the project frame, usage outside the project during the project time frame and usage after project completion.

B 4. Ethical Issues

Gender issues

The DILINET consortium is well aware of the recent efforts in Europe aimed at fostering equality for women and for ensuring equal opportunities for persons with disabilities[55] and other minorities, including linguistic. Although all of the partners in the DILINET consortium are equal opportunity employers, we are aware as a community that women are generally underrepresented in the scientific and engineering fields upon which the DILINET technology will repose. Some of the research domains covered by DILINET already have a relatively equal gender balance - such as linguistics and human factors, while in computer science and engineering the percentage of women is substantially lower. The project will encourage the involvement of more women experts in the project development.

One way in which DILINET can help remediate the situation is by having the senior women and minorities serve as role models, thus encouraging junior researchers and students to pursue careers in scientific and engineering fields. At the partner sites women currently represent between 15 and 40 percent of the research staff. We will actively encourage female doctoral students to carry out research related to the DILINET project, and some sites will offer financial support for such studies. We will also encourage women to visit the other partner sites to forge deeper relationships with other women working on similar research problems.

The research environments at the partner sites are conducive to flex-time, so that researchers can balance work and family constraints. We will try to organize consortium meetings so as to satisfy both work and family situations. The dates and host location will be chosen to facilitate travel and to the extent possible minimize disturbance to family life (avoiding school vacation, weekend meetings or travelling on weekends, minimizing the nights away from home).

While the technology developed in DILINET will on average perform equally well for users of both genders, there will inevitably be individual differences. Testing will be carried out in situations representing both genders.

To the extent possible, DILINET will follow the recommendations of the European Technology Assessment Network (ETAN) report on promoting gender equality in scientific research. The report, entitled "Science policies in the European Union: Promoting excellence through mainstreaming gender equality"[56].In particular, the report provides recommendations for good practice in equal opportunity recruitment, and to promote gender equality in decision making processes and dissemination activities.
The project management will:

- Adopt the appropriate measures encouraging women participation in the management of the project, in order to achieve a balanced consortium
- Support the implementation of the recommendations produced by the European Technology Assessment Network (ETAN) on the development and production of statistics and indicators, about the situation of women in scientific research

[55] W3C web content accessibility guidelines (WCAG)
[56] http://www.cordis.lu/etan/home.html

Project-wide approach to ethical issues

The major ethical issue in DILINET is to inform users about legal implications and ethical measures taken to ensure the protection of personal data as the date collection will take place throughout different activities (WPs) in the project. The partners involved in DILINET will prepare guidelines on ethical use of information for the project, describing how data will be collected, why and how it will be used. The guidelines will include concrete recommendations for the exploitation and integration of the research outcomes at the national level. In addition, the end user will receive specific information (End User License Agreement/Terms of Use) regarding the ethical aspects of the project. Upon request and aiming to maintain the highest level of transparency, the users will receive the executive summary of the project outcomes.

A special committee (Virtual Secretariat) will be established for all questions related to the legal and ethical issues, particularly privacy (linked to Task 4.3). The policy dealing with privacy issues will be aligned with the European directive guidelines[57] and participants/users will be informed about the objectives of the project and personal information will be held confidentially.

The fundamental principles outlined in various EU and UN international normative documents as human dignity, integrity of the person, the right to privacy - just to mention the most prevalent ones - will be fully respected and promoted within DILINET. All legal and ethical requirements of the member states where the project is implemented will be fulfilled.

Specific action with regards to ethical issues

As part of WP4 'Societal Issues', DILINET addresses specifically all legal, ethical and regulatory matters (T4.2) and all privacy protection and security concerns (T4.3) associated with such a project.

DILINET does not raise any specific Ethical issue. Should any ethical matter arise during the project life time, the project management will promote the assessment by a recognised ethics committee in consultation with EC.
All participants in DILINET will conform to the legislation and regulation in force in their respective countries. The concerned rules to be observed are:

- The charter of Fundamental rights of the EU
- Council directive 95/46/EC of Oct. 1995 on protection of individuals with regard to the processing of personal data and on the free movement of such data.

As a fundamental complement of the technical requirement specification, a thorough analysis and profiling of the legal and regulatory framework will be provided, including both i) privacy and data security requirements set forth at a European level, and ii) privacy laws in selected EU countries. These will not be static profiles but will also take into account the position held by local data protection Commissioners and actual practice of the selected countries (in some territories, actual practice differs from written law). A similar review of relevant law enforcement legislation will be undertaken to identify additional obligations and uniform technical models and standardised best practices. Rather than just provide an overview of all the rules in all the data protection laws and secondary rules and regulations, aim of this activity is to describe in a comparative and analytical way the laws in the selected Member States, in order to provide technical requirements which will ensure that the developed product will comply with the relevant rules of the EU regulatory regime. This will assist to get insights and solutions

[57] 2002/58/EC of the European Parliament and of the council concerning the processing personal data and the protection of Privacy in elecontronics communication sector; directive on privacy and electronic communications

to achieve a high level of integration between technical concepts and European laws and regulatory provisions.

Ethical issues table	YES	NO
Informed Consent		
Does the proposal involve children?		✓
Does the proposal involve patients or persons not able to give consent?		✓
Does the proposal involve adult healthy volunteers?		✓
Does the proposal involve Human Genetic Material?		✓
Does the proposal involve Human biological samples?		✓
Does the proposal involve Human data collection?		✓
Research on Human embryo/foetus		✓
Does the proposal involve Human Embryos?		✓
Does the proposal involve Human Foetal Tissue / Cells?		✓
Does the proposal involve Human Embryonic Stem Cells?		✓
Privacy		
Does the proposal involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)	YES	
Does the proposal involve tracking the location or observation of people?	YES	
Research on Animals		
Does the proposal involve research on animals?		✓
Are those animals transgenic small laboratory animals?		✓
Are those animals transgenic farm animals?		✓
Are those animals cloned farm animals?		✓
Are those animals non-human primates?		✓
Research Involving Developing Countries		
Use of local resources (genetic, animal, plant etc)		✓
Impact on local community		✓
Dual Use		
Research having direct military application		✓
Research having the potential for terrorist abuse		✓
ICT Implants		
Does the proposal involve clinical trials of ICT implants?		✓
I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	**NO**	

Explanation

The user-centric measurement of the project (WP6) potentially involves the processing of personal data and the tracking of the location of people. The first issue is the flip side of the unique ability of this type of measurement, namely to be able to measure social conduct in action (that is, language in real use). The second issue is inherent to the inclusion of mobile devices as they are the nexus of measurements.

We guarantee the privacy of our respondents by two complementary measures: (1) in the design of the measurement client and (2) validation by an independent third party.

Privacy issues may arise in the client for three different aspects: language determination, background information, and location information. We now discuss how we take privacy into account in the client's design, for each of these three dimensions.

In order to determine the language that is being used by the respondent, we send pieces of (potentially personal) information from the respondent's site to a central server. This involves sending a limited number of randomly taken nouns, adjectives or verbs over an encrypted connection. First, it is unlikely to derive from the centrally collected data any real personal data, since the data collection takes place at random. Second, once the algorithm at the server has identified the language the original input string is shredded.

During the first installation of the client respondents are asked to provide a limited number of general background characteristics. These background variables are the basis for analysis of the output data later on. To safeguard the privacy of the respondent, we use a randomly generated ID that is stored in a hash table. Given the sizeable number of IDs, even the use of rainbow tables will not enable reverse engineering of the IDs.

Finally, with regard to the location of the respondent, we will not collect or store the location information that is automatically being generated by the mobile device. In theory, though, should the background information store too fine grained geographical information, the combination of several background variables with data received could still identify a particular respondent. We counter this possibility by restricting geographic data to a very aggregate geographical location in the background variables (country), and broad categories that have at least several hundreds of respondents in one subclass (thus ensuring k-anonymity).

To ensure that all proper security measures are taken to safeguarded the privacy of our respondents, we have established a strict divide between the entity who will develop the user-centric client (Dialogic) and the entity (Kyos, a respected IT security firm) who will investigate the programming code, perform vulnerability tests, audit the security procedures that are taken at Dialogic and in the consortium as a whole, and provide advice on strong authentication and intrusion detection (at the server side).

In addition, WP 4 will develop the End User License Agreement/Terms of Use) for legal implications and inform the users about ethical measures to be taken to ensure the protection of personal data and the statistical usage of the collected data within the framework of the project. This work package will coordinate with all the partners guidelines on ethical use of information for the project describing how data will be collected, and why and how it will be used. The fundamental principles outlined in various EU and UN international normative documents as human dignity, integrity of the person, the right to privacy, etc. will be analyzed in order to be fully respected and promoted and feed task in the WP 4.

B 5. Annexes

B 5.1 Annex 1 – Letters of Intent

YAHOO!

Yahoo! Iberia, S.L.U.
Edifici IMAGINA
Avda. Diagonal, 177, 8th Floor
Barcelona
Spain

To Whom It May Concern

Letter of Intent: DILINET proposal to FP7 ICT Call 8
Objective ICT-2011.4.4 - Intelligent Information Management

Barcelona, 10th of January, 2012

Ricardo Baeza-Yates, representing Yahoo! Iberia, supports the DILINET proposal to be submitted on January 17th, 2012 to the FP7 ICT call 8 and commits to becoming the Chairman of the 'Scientific External Advisory Board' as soon as the project starts and for its full duration.

I intend to coordinate the activity of the Board to closely follow and evaluate the progress of this project, and to actively participate in the two planned assessment and evaluation workshops as Chairman of the DILINET Scientific External Advisory Board.

Furthermore, I will explore possibilities to experimentally use the reference applications under development whenever appropriate and to provide feedback with regards to their relevance and potential usefulness.

With best regards,

YAHOO!
María de Molina, 40 - 5ª
28006 Madrid
Tel.: 91 411 87 00 / Fax: 91 411 88 48

Dr. Ricardo Baeza-Yates
VP of Research for EMEA & LatAm
Yahoo! Research Barcelona

Yahoo Iberia S.L.U.
Avda. Diagonal 177, 8ª pl
08018 Barcelona
RM Madrid. Tomo 14.219, Folio 173, Sección 8, Hoja M-234.114, Inscripción 1/A.
C.I.F. B/61.710.737

MENON Network
Rue des Deux Eglises 35
1000 Bruxelles Belgium

To Whom It May Concern

Letter of Intent: DILINET proposal to FP7 ICT Call 8
Objective ICT-2011.4.4 - Intelligent Information Management

Brussels, 9/01/12

Fabio Nascimbeni, representing the MENON Network, supports the DILINET proposal to be submitted on January 17th, 2012 to the FP7 ICT call 8 and commits to becoming a member of the 'Societal Matter External Advisory Board' as soon as the project starts and for its full duration.

I intend to closely follow and evaluate the progress of this project, and to actively participate in the two planned assessment and evaluation workshops as member of the DILINET Societal Matter External Advisory Board.

Furthermore, I will explore possibilities to experimentally use the reference applications under development whenever appropriate and to provide feedback with regards to their relevance and potential usefulness.

With best regards,

Fabio Nascimbeni
Director, MENON Network

ORGANISATION INTERNATIONALE DE la francophonie

N/Réf : DLC/DFN/20120113-014

Objet : Lerttre d'intention en faveur du projet DILINET Paris, le 13 janvier 2012

Par la présente l'OIF représentée par son Directeur de la Direction de la langue et de la Culture, Monsieur Frédéric Bouilleux, et son Directeur de la Francophonie numérique, Monsieur Pierre Ouédraogo, confirme son intérêt et engagement envers le projet DILINET dans les termes ci-après.

L'Organisation internationale de la Francophonie se préoccupe de la place de la langue française et des nombreuses langues partenaires des 75 Etats membres ou observateurs, dans le monde en général et dans le cyberespace en particulier. Elle maintient un Observatoire de la langue française dans le monde et publie avec régularité les données de cet observatoire. C'est dans ce contexte que l'OIF s'est intéressée à l'initiative de l'organisation MAAYA pour créer un consortium de recherches avancées susceptibles de dépasser les limites actuelles pour la production d'indicateurs des langues dans l'Internet et d'aider à la compréhension des évolutions de la structure de la Toile et des moteurs de recherche qui permettent d'y appréhender des contenus en évolution permanente. À travers sa Direction de la Francophonie numérique, l'OIF a participé financièrement, aux côtés d'autres organisations internationales, aux pré-études conduites par MAAYA et qui ont abouti à la définition du projet DILINET. L'OIF souhaite continuer à apporter sa contribution, intellectuelle cette fois, et participer au suivi de cette opération. L'application prévue et correspondante à la Tâche 11.3 pour un pilote de politique publique pour les langues, à la définition de laquelle nous avons contribué, nous intéresse plus particulièrement, ainsi que l'ensemble des activités prévues pour l'exploitation et la dissémination.

Conscient de l'importance de la représentation compétente des utilisateurs des produits des recherches du projet DILINET, nous acceptons de participer pleinement au Conseil de Surveillance sur les aspects sociétaux que ce projet entend constituer, avec des fonctions non rémunérées de conseil et assistance là où nos compétences pourront apporter et dans un rôle d'évaluation de certains des produits prévus par le projet.

Nous acceptons que cette lettre d'intention soit jointe à la proposition du projet DILINET et continuerons d'œuvrer pour le succès de cette démarche qui nous parait déterminante pour le futur des indicateurs de la société de l'information et pour une meilleure caractérisation de la Toile.

Frédéric BOUILLEUX
Directeur de la Langue française
Et de la Diversité culturelle et linguistique

ORGANISATION INTERNATIONALE DE LA FRANCOPHONIE
OIF
19-21 avenue Bosquet - 75007 PARIS

Pierre OUEDRAOGO
Directeur de la Francophonie
numérique

ORGANISATION INTERNATIONALE DE LA FRANCOPHONIE
19-21, AVENUE BOSQUET - 75007 PARIS (FRANCE)
TÉL. +33 (0)1 4437 33 00 - TÉLÉCOPIE +33 (0)1 44 11 12 76
www.francophonie.org

Language Observatory Project

c/o Nagaoka University of Technology
1603-1 Kamitomioka-machi,
Nagaoka, Niigata, Japan

To Whom It May Concern

Letter of Intent: DILINET proposal to FP7 ICT Call 8
Objective ICT-2011.4.4 - Intelligent Information Management

13/01/2012

Yoshiki Mikami, representing Language Observatory Project, supports the DILINET proposal to be submitted on January 17th, 2012 to the FP7 ICT call 8 and commits to becoming a member of the 'Scientific External Advisory Board' as soon as the project starts and for its full duration.

I intend to closely follow and evaluate the progress of this project, and to actively participate in the two planned assessment and evaluation workshops as member of the DILINET Scientific External Advisory Board.

Furthermore, I will explore possibilities to experimentally use the reference applications under development whenever appropriate and to provide feedback with regards to their relevance and potential usefulness.

With best regards,

Yoshiki Mikami

Leader of Language Observatory Project,
Professor Nagaoka University of Technology